

Strength through diversity

A plan to reform appropriations committees in the US Congress — and create one devoted to science — is unlikely to come to fruition. Which is just as well.

The research and development programmes of the US federal government are scattered across departments and congressional committees in a rather haphazard way. For decades, tidy minds have suggested improvements. But these have been ignored — and US science has gone from strength to strength.

The latest proposal comes from Tom DeLay (Republican, Texas), the majority leader in the House of Representatives, who wants to streamline the appropriations subcommittees that set budgets. His plan reportedly includes a subcommittee for science that would take under its wing NASA, the National Institutes of Health (NIH), the National Science Foundation (NSF) and the research programmes in the Department of Energy (DOE).

At present, the science agencies are scattered across different appropriations subcommittees, so NASA competes directly with the veterans affairs and housing departments for money. The NIH competes with agencies such as the education department. This may lack logic, but it has served science well. The scientific agencies have each found champions in both houses of Congress who know what they do and support their work. The current system also means that the major science agencies don't compete directly with each other for money — decisions about how to fund space science are made separately from choices on biomedical research. The system has put a brake on the sudden diversion of money into spheres that are fashionable or politically correct, and has helped to build the strongest and most diverse research enterprise the world has ever seen.

Putting everything in the hands of a single appropriations subcommittee could unravel all that. Agencies would compete bitterly

for funds and the wrong priorities might take hold. The dynamics of congressional committees would suggest that members with large NASA or DOE facilities in their districts would come to dominate the subcommittee, draining clout from agencies such as the NIH and the NSF, which give small grants to researchers all over the country.

Maybe that's part of what DeLay is seeking. He recently intervened to restore the NASA budget — partly in defence of President Bush's human spaceflight plans, and partly to flex some muscle on behalf of the Johnson Space Center in his own district. He may believe that a research subcommittee would better serve NASA's interests. But the broader intention of his plan, according to Capitol Hill newspaper *Roll Call*, which broke the news on 6 December, is to realign the budget system to help the Republican leadership to shrink the federal government.

But even the man known in Washington as 'The Hammer' will have his hands full reforming such an entrenched system. The House will be reluctant to break up its current structure, and the Senate is expected to be hostile. No one expects any action soon, but that could change if the Republicans strengthen their grip on the Senate in 2006.

This inertia is probably good news for science, and for science lobbyists, who complain about the current system but know how to work it. As one biomedical lobbyist said last week: "I would rather fight nine small battles for adequate funding for research than have one large budget battle, in which a gain for one area of science came at the expense of all the rest." Excessive concentration of power over distributing research funds would lead to more political interference, more funding instability, and less diversity in US research. ■

Fishing's secretive controllers

Those who are publicly funded to regulate harvesting of the oceans should stop barring the public from their discussions.

As the international agency controlling fishing for 30 migratory species in the Atlantic Ocean, the International Commission for the Conservation of Atlantic Tunas (ICCAT) has a vital ecological role. From the spawning waters of the Gulf of Mexico, up across the feeding grounds of the North Atlantic and into stock-rich regions of the Mediterranean, ICCAT decisions on allowable commercial catches can have a huge environmental impact. They affect target species, such as bluefin tuna or swordfish, creatures ensnared as bycatch — notably sharks or dolphins — and their food webs.

Timely, proper assessments of the stocks of specific species, analysis of the age and size of captured fish, and understanding of the effect of years of overfishing should be integral components of ICCAT deliberations. These should lead to science-based management decisions on the economically important and culturally rich resources under the stewardship of the Madrid-based agency. They don't.

Disturbingly, the media and the public are blocked from observing ICCAT discussions. Only in recent years did ICCAT allow some non-governmental organizations to attend its conferences, but even they must pay an 'observer fee' of at least US\$500, a move clearly designed to limit access.

Officials at the US National Marine Fisheries estimate that US taxpayers contributed some \$250,000 to host ICCAT's meeting in New Orleans last month, which included representatives of 23 non-member fishing nations. Deliberations — funded by the taxpayers of the respective nations — affect the international public's resources. Given this public funding, such meetings should be open, but, apart from a press conference at the start, reporters were barred.

What do ICCAT members have to hide? Shamefully, representatives of the major markets in Asia and the heaviest fishing nations in Europe choose to meet in private and put their economic interests ahead of science-based management. This is what happened in New Orleans (see *Nature* **432**, 539; 2004).

One ICCAT delegate suggested that the reason the press was not allowed in was because this would disturb developing nations. Excuses like this show the lengths to which officials will go to keep proceedings under wraps.

Transparency is essential if we are to understand what is happening to ocean resources. The Convention on Biological Diversity operates in such a manner. ICCAT should follow its example, starting by opening up its next meeting in Japan in April. ■

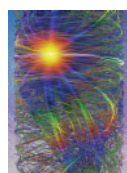




Just say 'non'
Staff resist relocation
plans for Pasteur
Institute
p788



Locked out
Inquiry stalls as
Pentagon denies
access to data
p789



Small fry
Nanotechnology rush
raises quality-control
questions
p791



Selected few
Bush names his
nominees for
energy and health
p792

NASA administrator quits post to launch career as academic

Tony Reichhardt, Washington

The future of the US space programme faces fresh uncertainty this week after Sean O'Keefe quit as administrator of NASA to take up a lucrative academic post.

O'Keefe resigned on 13 December to pursue a job as chancellor of Louisiana State University in Baton Rouge, which had aggressively recruited him and offered a reported \$500,000 salary. The move came as a surprise to many in the space community, who had hoped that, before he left, O'Keefe would clear up some of the more pressing problems that arose during his tenure.

A former Secretary of the Navy, O'Keefe was widely credited with addressing cost overruns on the International Space Station after his appointment in December 2001 — albeit by eroding the station's technical capacity. But after the loss of the space shuttle Columbia in February 2003, O'Keefe became an enthusiastic cheerleader for President George Bush's ambitious 'Vision for Space Exploration', including a return to the Moon.



In an important victory for O'Keefe and the White House, NASA last month obtained nearly its full 2005 budget request of \$16.2 billion from Congress. Yet the future of the 'vision' is far from assured, space analysts say, and NASA's financial troubles are by no means over. The shuttle has so far cost \$750 million more to fix than was anticipated. And a public dispute over how to save the ailing Hubble Space Telescope — which may need as much as \$2 billion that NASA doesn't have in its budget — has been a mounting headache for O'Keefe.

A National Academy of Sciences panel released a report on 8 December that directly challenged the administrator by calling for astronauts to repair the telescope as originally planned. A proposed robotic servicing mission would be too risky in the near term, said the panel, and sending astronauts to Hubble is not significantly more dangerous than sending them to the space station — despite O'Keefe's protestations to the contrary. The administrator made no official response to



Hubble trouble: working out a fix for the telescope has been a headache for Sean O'Keefe.

the report, other than to say that NASA is still studying the robotic servicing option, which comes up for a technical review in March.

Many observers now see O'Keefe's insistence against sending astronauts to fix Hubble as a significant blunder. It underestimated public support for the telescope, they say, and alienated influential politicians such as Senator Barbara Mikulski (Democrat, Maryland), whose state is home to Hubble's control centre.

Taking flight

Although many scientists respected O'Keefe's management style and sympathized with his attempt to give the astronaut programme a fresh direction, the potential dominance of the Moon mission and the Hubble controversy made others fret that science had lost its pre-eminence at NASA. Space science had been largely insulated from budgetary problems affecting the space shuttle and station in the past decade. Now, says Jonathan Lunine, a planetary scientist at the University of Arizona in Tucson and a frequent member of NASA advisory committees, "some of the virtual firewalls between space science

and manned spaceflight are breaking down".

O'Keefe also began another, less heralded transformation at NASA by filling many top agency positions with former military officials, many of whom had no previous space experience. Jobs ranging from chief financial officer and general counsel to the head of the new exploration programme — retired Navy Rear Admiral Craig Steidle — have gone to people with Pentagon backgrounds. This partly reflects O'Keefe's work experience, but also signals the Bush administration's interest in fostering a closer relationship between military and civilian space programmes.

The use of military expertise extends to outside advisers, such as retired Air Force General Lester Lyles, a former head of the US Missile Defense Agency, who has been tapped to head NASA's oversight committee for the Moon mission. In a speech to the Air Force Association last month, Lyles praised O'Keefe for "getting to an organization that looks very much more like the Department of Defense".

That trend may continue after O'Keefe's departure. Rumoured successors include several military men, among them retired Air Force General Ronald Kadish, who recently led the US missile defence programme. ■

Tempers flare over plan to move Pasteur labs

Declan Butler, Paris

Scientists at the Pasteur Institute in central Paris are protesting over plans to relocate part of the prestigious research centre to a commercial zone on the outskirts of the city. So bitter has the row with management become that the institute has called in an outside mediator to defuse the tension.

Over the past few months, hundreds of Pasteur scientists have received letters from management informing them that their labs will be moving to Fresnes for several years while the central campus is renovated. Some of the institute's most senior scientists are among those told they will move. Staff object not only to the move itself but also to how the decisions have been made.

Researchers say that the 17,000-square-metre Fresnes site in the Val-de-Marne region southeast of Paris is in an undesirable commercial zone with poor public transport. And with little hope of attracting good young scientists, they argue, the move would be a recipe for intellectual death by isolation.

Almost half of the staff have already signed a petition stating that they "understand neither the necessity, nor the rationale" of the move, which they claim "would be a major scientific handicap" to the institute.

The Fresnes site, estimated to be worth some €13 million (US\$17 million), was donated to the Pasteur earlier this year by the drug firm Pfizer. The institute says the site solves space problems for the renovation and will allow for a planned expansion. Local authorities have promised to provide subsidies of €4.3 million for the Pasteur as the core of a planned biotech science park, according to the Val-de-Marne Economic Development Agency.

Agnès Labigne, head of the Pasteur's



Researchers at the Pasteur Institute claim a move to the outskirts of Paris will compromise their work.

Pathogenesis of Mucosal Bacteria unit, says the renovations are being used as a "pretext" to justify "Fresnes at any price". She stepped down as head of her department in March because, she says, the Pasteur management ignored the conclusions of a task force that the renovation could be carried out without displacing groups to Fresnes.

The relocation row has also brought to a head tensions over what staff see as the brash management style of the institute's director-general, Philippe Kourilsky. He wrote in a letter to the scientists that the early notices about the move were "badly received by many of you, who judged it brutal, and were rightly astonished that it was not preceded or accompanied by personal contact with management". He says management had intended to talk to them on the day the notes were sent.

Kourilsky was travelling and unavailable for comment. But senior vice-president for

scientific affairs Stewart Cole, acting spokesman in Kourilsky's absence, admits that there is a problem. "At the moment there are a lot of difficult decisions to be made, or to be reversed. I don't think the consultation has been good," says Cole. But he adds: "Everyone genuinely feels they are acting in the best interests of the Pasteur."

Cole and Philippe Sansonetti, head of the institute's cell-biology department, have now persuaded management to bring in a mediator, John Skehel. Skehel is director of Britain's National Institute for Medical Research in London — itself no stranger to fights over site relocation (see *Nature* 432, 662; 2004).

"The situation has got completely out of hand," says Cole. "We need a few months to see what the most sensible way out is. But we need to do this in a concerted, open and constructive manner." Skehel is expected to deliver his initial conclusions by the end of January. ■

Relaxed rules open path to genomic data on disease

David Cyranoski, Washington

To the delight of researchers mining the human genome for disease-related genes, restrictions on the use of data from the International HapMap Project were dropped last week. The information can now be incorporated into other databases, allowing easy access for scientists around the globe.

Researchers seeking the genetic component of a specific disease are interested in single nucleotide polymorphisms (SNPs) — points in the genome where one base often differs between individuals. But there are an estimated ten million common SNPs, so it is quicker to search for common haplotypes linked to a disease — larger blocks of DNA that are often inherited together. The

HapMap project is a consortium of some 25 groups, which have been collecting data on haplotypes since October 2002.

When the US\$100-million project began, organizers were worried that people would mine the data and attempt to patent haplotypes. So users had to agree not to limit access to the information, for example by patenting it, and not to share the data with anyone who hadn't made the same agreement.

But this had the unfortunate side effect that data could not be integrated into large, open-access genomic databases such as Ensembl. "We could pick up the data and do a lot of analysis, but we couldn't give it back to anyone," says bioinformatician Ewan Birney, who heads Ensembl.

Those restrictions have now been lifted. Birney predicts that the increase in Ensembl data will throw up haplotypes for diseases such as Alzheimer's, diabetes or heart disease within the next two years. "It will switch on the light," he says. "For some diseases, major risk factors will be discovered."

HapMap partners hope that haplotypes related to specific diseases will be commercialized, says Lisa Brooks, a director of the US Genetic Variation Program, but that basic haplotype information will remain freely available. Brooks adds that enough data and analysis are now public that claims of novelty needed to secure a patent should be difficult to prove. ■

♦ www.hapmap.org



Results of missile defence tests conducted by the MIT Lincoln Laboratory will not be open to scrutiny.

Pentagon blocks MIT inquiry into missile data fraud claims

Geoff Brumfiel, Boston

The Massachusetts Institute of Technology (MIT) has been forced to abandon a fraud inquiry at one of its laboratories after the Pentagon denied it access to the suspect data.

The university wanted to investigate contentious missile defence tests that took place at its Lincoln Laboratory six years ago. "Without access, the investigation cannot be conducted," Charles Vest, MIT's outgoing president, said in a statement on 1 December. Vest left his position on 5 December.

Pentagon officials say that the test data are classified and cannot be released to an investigatory panel on the grounds of national security. But critics see the Department of Defense stance as a political attempt to block further inquiry into the research.

Some government watchdogs say that the Pentagon's approach threatens to undermine the integrity of academic institutions that, like MIT, conduct classified research. "It's an extraordinary situation," says Steven Aftergood, who oversees a project aimed at reducing government secrecy, at the Federation of American Scientists in Washington DC. "It should prompt a rethink of universities' policies on the subject."

At the heart of the debate is a long-disputed series of tests of a sensor designed to detect incoming missiles. Shortly after the tests were conducted in 1997 and 1998, a former engineer from the US defence contractor responsible for the sensor came forward with documents that she claimed proved the contractor had tampered with data to hide the sensor's failure.

An investigation carried out by the General Accounting Office backed the engineer, but a subsequent investigation by the Lincoln Laboratory itself, for the federal government,

found that no data had been tampered with.

That led Ted Postol, an MIT physicist and established critic of missile defence, to demand that university officials investigate whether the laboratory had acted improperly. "Lincoln Laboratory knew that the sensor had failed," he contends. "And they failed to tell federal agents."

Postol wrote to Vest in 2001 asking for a separate university-led inquiry. That inquiry was launched in spring 2002 by Edward Crawley, an aerospace engineer at the university, and by the end of that year, Crawley had found enough evidence to justify a full investigation. The plan was for a review to be conducted by a select group of academics from outside MIT who had the security clearance needed to view the classified data.

But the investigation was never begun because it was opposed by the Pentagon's Missile Defense Agency. Agency officials declined to comment, but said in a statement that the agency had already been exonerated by earlier inquiries, and that a new investigation only risked leaking classified information about the missile defence system. "The extreme sensitivity of the information at issue precluded granting MIT's request," the statement says.

"The implications of this are very serious," says Sheila Widnall, another MIT aerospace engineer and former secretary of the US Air Force. Widnall chaired a panel to review the university's policy for managing classified research. For such research to be worth the trouble, the panel concluded in 2002, it must be conducted with the highest standards of integrity and in a way that is independent of the funder. "What's happening now does not meet that standard," she says.

Absence of promised bioterror committee worries researchers

Erika Check, Washington

Researchers are voicing concern that a government initiative, designed to help biologists ensure that their work isn't misused by bioterrorists, isn't getting off the ground.

Nine months after the US government said it would create a powerful committee to advise it on issues related to bioterrorism, the committee hasn't been appointed, never mind held a meeting.

It was 4 March when Tommy Thompson, then health secretary, announced the creation of the National Science Advisory Board for Biosecurity (NSABB). It was expected to advise the government on tough questions about overseeing 'dual-use' research, which could have civilian or military uses, and to write a code of conduct for scientists.

But the board hasn't appointed either members or support staff, alarming senior scientists who lauded its creation. "This is an important initiative, and it needs to move forward," says Ronald Atlas, co-director of the Center for Deterrence of Biowarfare and Bioterrorism at the University of Louisville in Kentucky and former president of the American Society for Microbiology. "Without the NSABB, the scientific community remains in limbo and guessing about what steps will be constructive as we try to prevent the misuse" of science, he adds.

Officials at the National Institutes of Health are coordinating the formation of the board. "We are close to making an announcement," says Bill Hall, a spokesman for the health department. "It's not going to be Monday, but it's not going to be six months from now."

Some biosecurity specialists worry that if the board isn't up and running soon, the government could bypass it and take extreme steps to regulate some types of dual-use biology. They also think the board is needed now to help oversee the massive biodefence research programmes set up by the United States since the terrorist attacks of 11 September 2001.

"There are so many expectations heaped on this board that they probably never would have met them," says one biosecurity specialist. "But it's hard to meet them if you don't even convene."

For more news and analysis go to
news@nature.com
www.nature.com/news

Oil refinery site may be home to ancient relics

Rex Dalton, Yuma

Plans to build the first oil refinery in the United States for nearly 30 years are heading for trouble over claims that the intended site may harbour important Native American archaeology.

The site is in the Gila River valley in Yuma county, Arizona, and is part of 22,600 hectares currently owned by the US Bureau of Reclamation. The region served as a major route across the Sonora Desert for centuries before the railroads were built.

Archaeological surveys have identified lots of finds in the valley — from rock paintings to former dwelling locations — that might help answer intriguing questions about life in an environmentally lush desert passage. “Archaeologists have long been interested in the valley,” says Jeffrey Altschul, an archaeologist at Statistical Research, a Tucson-based consultancy.

The reclamation bureau is in the process of transferring the 22,600 hectares to the Wellton-Mohawk Irrigation and Drainage District. The district then plans to sell 1,200 hectares of it to Phoenix-based Arizona Clean Fuels, for construction of the \$2.5-billion oil refinery.

No archaeological sites have been identified on the proposed site for the refinery, officials say, because it hasn’t been surveyed yet.

As required by law, the bureau hires archaeologists to examine public land that is going to be sold or transferred, to ensure that important sites are charted for protection. But the irrigation district has pushed to



Building site: the Gila River valley could house the first US oil refinery to be built for decades.



speed up the deal before archaeological studies are complete so that it can sell the land for the refinery.

Given the high financial stakes and Arizona’s history of building over archaeological sites, community watchdog groups and Native American tribes are expressing concerns about the plan. “People need to take a hard look at what is going on,” says Steve Brittle, who runs Don’t Waste Arizona, a Phoenix-based environmental group.

Brittle and other critics fear that the bureau may buckle under commercial and political pressure, and transfer the land too quickly. Bureau officials in Yuma and Washington say that they are following all legal requirements, and will continue to do so.

The handling of the site could set a precedent for other sell-offs, as the bureau seeks to dispose of several sites in the western United States. The Gila River transfer is one of about 16 currently being undertaken by the bureau — and the first known to contain important archaeological remains.

Archaeological studies are now just getting under way on a larger land transfer of 33,200 hectares of the Humboldt River basin in northern Nevada, for example. Initial analysis indicates that these lands may also include significant archaeological sites.

The bureau will release a more comprehensive analysis of the Arizona site early next year. Native American tribes pushed for a more in-depth investigation after claiming that a 2002 government study wasn’t sufficiently comprehensive.

But irrigation-board officials are becoming frustrated with the length of time taken to do the studies, and they are considering giving up on their plans for some of the archaeologically sensitive lands, so that they can secure the sale of the refinery tract.

This week in Yuma, the county government was expected to take the first step towards granting permission for the refinery project, which is also expected to draw opposition on environmental grounds. ■

Frenchman is most thanked computer scientist

Declan Butler, Paris

Who is Olivier Danvy? He is the most thanked person in computer science, according to an analysis of acknowledgements on nearly a third-of-a-million scientific papers.

The analysis, published this week, marks the debut of text-mining software from the lab of Lee Giles, a computer scientist at Pennsylvania State University.

The software opens up a largely untapped mine of information about how scientists and agencies affect all fields of science, says Jon Kleinberg, a computer scientist at Cornell University in Ithaca, New York. “Something that once seemed like it would require an enormous amount of manual labour has suddenly become feasible,” he says.

Giles used the program to extract data on who had thanked whom in 335,000 papers in his lab’s CiteSeer archive of computer-science papers (C. L. Giles and

I. G. Councill *Proc. Natl Acad. Sci. USA* 101, 17599–17604; 2004). The results reveal the people and funding agencies who received the most thanks in computer science.

Research on acknowledgements has been hampered by the lack of a data repository to study, says Kleinberg, contrasting this with citation statistics, which have been compiled manually for more than four decades by Thomson ISI in Philadelphia.

Acknowledgements are more arbitrary than citations. “Peer review always requires you to incorporate the latest and relevant citations. But no one is checking whether all contributors are referenced in the acknowledgements,” says Erik van Mulligen, chief technology officer at Collexis, a text-mining company in Geldermalsen, the Netherlands.

Social scientists have already looked into the problem. They have broken acknowledgements into six main types,

including support from funders, and the ‘conceptual’ support provided by scientists such as Danvy.

Eugene Garfield, founder of ISI, agrees that the various types of acknowledgement need to be categorized with care. “Otherwise you’ll get quite a mishmash,” he says.

Danvy, a French researcher who works on programming languages at the University of Aarhus in Denmark, says he was “stunned to find my name at the top of the list”. After reflection, he put it down to a “series of coincidences” — he is multidisciplinary, well travelled, runs an international PhD programme, is a networker and belongs to a university department with a long tradition of having many international visitors.

“It’s a snowball effect,” says Danvy, who admits to being a helpful sort of fellow. “I encourage people a lot, and I advise many students on their papers.” ■

Growing nanotech trade hit by questions over quality

Jim Giles, London

The surge of interest in nanotechnology has created buyers and sellers of materials who aren't properly equipped to do business, analysts have warned.

A survey of nanotechnology companies has revealed a series of problems with the supply of nanomaterials — tiny tubes, wires and cages with dimensions of just billionths of a metre. Batches may be contaminated or the nanoparticles are improperly shipped and clump together during delivery. "We heard one horror story after another," says Matthew Nordan of Lux Research, the New York-based consulting company that published the survey on 8 December.

One semiconductor company found that almost one-third of a sample of carbon nanotubes it had purchased consisted of iron left over from the production process. "Not only is the iron not what the lab paid for, it's a contaminant that would never be allowed in a semiconductor fabrication facility," notes the report, which drew on interviews with more than 100 managers.

Nordan says that the recent explosion in research funding for nanotechnology is partly to blame. In the United States, government spending in the field has increased sixfold since 1997 and now stands at almost \$1 billion a year. Nordan and other industry experts suggest that customers and suppliers have rushed into the field, and have yet to develop good systems for working with each other.

Some of the worst complaints concerned university laboratories that sell nanomaterials as a side business. One buyer complained to Nordan that "every professor who wants a beach house, and who works on nano-something, has started a company".

But customers also come in for criticism. Peter Eklund, a nanotechnology researcher at Pennsylvania State University in University Park, is a founder of CarboLex, a nanomaterials supplier based in Lexington, Kentucky. Eklund says he has received only two or three complaints in seven years of business, and all have related to poor handling by researchers. One customer complained that the tubes were shoddy, for



Big business? Interest in nanomaterials, such as this nanowire, has driven a rapid rise in supply and demand.

example — but after discussing their research, Eklund found that the researcher had inadvertently vaporized the tubes while attempting to purify them.

"There is a lack of communication between manufacturers and customers," agrees Mark Banash, the senior engineer at Zyvex, a nanotechnology firm in Richardson, Texas. Banash is the force behind one potential solution to the problem: he runs a certification scheme, under which suppliers are approved if they reach an agreement with Zyvex about the characteristics of the nanotubes they sell and their ability to supply them on time. One supplier has already signed up and Banash says two more should be approved in the next few months. Although the scheme was designed for Zyvex's needs, Banash hopes it will also help other firms to choose suppliers. ■

Researchers call for more materials science in school

David Cyranoski, Boston

Materials scientists in the United States are mounting a concerted effort to get their discipline taught at high school.

"It's time to stop looking at science as just chemistry, physics or biology. We need to recognize materials science as fundamental," says Michael Rubner, a polymer scientist at the Massachusetts Institute of Technology. "The basic notion of what a materials scientist does should be passed on at the earliest possible time."

Teachers involved in the effort admit that materials science is an unfamiliar concept to their pupils. "I don't think anyone's heard of it," says David Ruth, who teaches at South Seneca High School in Ovid, New York. But Ruth believes that his efforts to bring materials to life are drawing kids in.

Thomas Stoebe, a materials scientist at the University of Washington in Seattle, says the discipline is already reaching schools as a result of week-long teacher-training programmes he has helped to establish. More than 200 teachers across the country have taken the lessons back to their classrooms.

Last month the National Science Foundation brought 90 teachers to a Materials Research Society meeting in Boston for a symposium on teaching materials science in schools. It has also been expanding its Research Experience for Teachers (RET) programme, which sponsors about 140 teachers each summer to train at materials-science departments in leading universities.

RET programmes have helped teachers develop ways of introducing such topics into the classroom. One lesson, developed at Pennsylvania State University, focuses on the question: "How do planes fly?" It includes the construction of a wind tunnel to teach basic concepts in aerodynamics and materials.

But some teachers worry that the No Child Left Behind act, which calls for standardized testing on core subjects, will squeeze out more specialized teaching. "With this programme, we're actually worse off," says Caroline Goode of the National Science Teachers Association in Arlington, Virginia.

Even though programmes that train just a few hundred teachers receive little funding, Rubner says they can still have "an enormous impact" as teachers return from their training programmes and spread the word to others. "It is having a ripple effect," he says. ■

Libya suggests it will drop death penalty for Bulgarian nurses

London Five Bulgarian nurses and a Palestinian doctor who were sentenced to death by a Libyan court in May, for allegedly infecting 400 children with HIV, may be spared. After intense pressure by the international community, including protests by scientists, the son of Libyan leader Muammar Gaddafi said last week that the government would commute the sentences.

"I rule out the possibility of executing the Bulgarian defendants," Seif al-Islam Gaddafi told The Associated Press. He added that Libya, which is seeking rehabilitation by the international community after decades of ostracism, will reconsider its use of capital punishment so that it will apply only in "limited, narrow cases". Seif Gaddafi has no official government position, but often speaks on behalf of his father.

The health workers have been in prison for five years (see *Nature* 430, 277; 2004), even though scientists say that many of the children were infected as early as 1994, long before the workers arrived in Libya. The evidence suggests that the infections came from another patient via dirty syringes.

Bush names secretaries for energy and health

Washington Candidates have been selected to run the US departments of energy and health. On 10 December, President George W. Bush nominated Samuel Bodman, deputy secretary of the commerce department, as the next energy secretary. And on 13 December he selected Mike Leavitt, head of the Environmental Protection Agency, to lead the health department.

Touted by Bush as "a problem solver", Bodman, a former chemical engineer, would oversee much of US physics research and the nuclear-weapons programme. If confirmed by the Senate, he will succeed Spencer Abraham, who stepped down last month.



In the frame: President Bush nominates Samuel Bodman (left) as energy secretary.

Gates cash lifts hope of cheaper malaria drug

San Francisco Research into an effective antimalarial drug received a boost on 13 December with the announcement of a US\$43-million grant from the Seattle-based Bill & Melinda Gates Foundation.

The grant will support the development of a synthetic form of artemisinin, the drug of choice for treating malaria. Artemisinin is derived from a plant (right) grown in China and Vietnam. Increasing demand has far outstripped supply, leading to a crisis in countries that have made treatment with the drug part of a national strategy (see *Nature* 432, 259; 2004).

The new funding will go to the Institute for OneWorld Health in San Francisco, a non-profit drug company, which will work in partnership with Amyris Biotechnologies and chemical engineer Jay Keasling of the University of California, Berkeley. Keasling is developing a method for turning genetically modified bacteria



into artemisinin factories (see *Nature Biotechnol.* 21, 796–802; 2003).

A spokesperson for the Institute for OneWorld Health said that the process should bring down the price of the drug from a prohibitive \$2.40 per adult course to less than \$1.

Leavitt, a former governor of Utah, would replace Tommy Thompson, who resigned on 3 December. He would be responsible for all US biomedical research and public-health agencies, including the National Institutes of Health.

US company buys into China's biotech market

Washington US biotech firm Invitrogen made a big move into China last week with the US\$8-million purchase of Shanghai-based Bio Asia, a major reagent manufacturer and distributor. The purchase was announced just days before World Trade Organization (WTO) guidelines opened the Chinese biotech industry to foreign-owned companies.

Until the change in the WTO's rules, "the playing field was heavily stacked in favour of local companies", says Ian Chapman-Banks, Invitrogen's Asia Pacific marketing director. "We were at the mercy of our distributors."

Usually reagents are distributed by local companies, often causing delays and frustration among Chinese researchers, says Chapman-Banks. But over the next two years, distribution should be shortened to a matter of days, he says.

Europe's biology club gives Iceland a warm welcome

London On New Year's day, Iceland will become the eighteenth state to join the European Molecular Biology Laboratory (EMBL). EMBL, which has its main laboratory in Heidelberg, Germany, provides shared facilities and expertise that encourages international collaboration between its 1,300 staff members.

Despite its small population, Iceland has made "significant contributions in the field of molecular-biology genetics", says Fotis Kafatos, EMBL's director-general.

"Since the mid 1980s Iceland has invested quite considerably in biotechnology," says Thorgerdur Katrin Gunnarsdottir, Iceland's science minister. For example, Icelandic company deCODE Genetics has combined the country's unusually detailed genealogical records with DNA sequence data to hunt for disease-related genes in the population.

Google finds itself in court in trademark row

Washington The American Chemical Society (ACS) has sued search-engine company Google in a US district court for trademark infringement and unfair competition.

The Washington-based ACS has been selling a search engine and database for chemistry-related data called SciFinder Scholar since 1998. It was indignant last month when Google, based in Mountain View, California, introduced a free scientific search engine called Google Scholar (see *Nature* 432, 423; 2004).

The ACS says that scientists often shorten the name of its research tool to 'Scholar', so there could be confusion with Google's product. Google has said that it will continue to use the name, but has not formally replied to the complaint.

"We feel we have no choice but to do this," says Flint Lewis, general counsel for the ACS. "We are protecting the reputation and goodwill of a service that's been around for six years, and data that we've built up for 90 years."

♦ <http://scholar.google.com>

♦ <http://www.cas.org/SCIFINDER/SCHOLAR>



M. PROBST/AP

Time lord: Fuat Sezgin surveys a globe reconstructed using ancient Arabic texts.

Rebuilding the past

Western science owes much to Islam's golden age — a debt that is often forgotten. To help redress the balance, Fuat Sezgin has reconstructed a host of scientific treasures using ancient Arabic texts. Alison Abbott reports.

Few passers-by in this leafy German suburb realize that the wonders of medieval Islam are just a few steps away. But those who enter Frankfurt's Institute for the History of Arabic-Islamic Science find themselves among the wonders of the historic Middle East. Fuat Sezgin, professor emeritus of the history of science at the University of Frankfurt, guides his guests through a labyrinth of tiled and mirrored rooms to the white-walled chambers of an exquisite yet little-known museum.

And there it is: an Aladdin's cave of science treasures. Each one has been recreated, thanks to Sezgin's labours, from descriptions in ancient texts of the Arab world. Some 800 newly built instruments — from ornate astrolabes to complex water clocks — are on display within the museum's 13 rooms.

Sezgin's long academic life has been quietly dedicated to retrieving Arabic science history. Now, with his eightieth birthday just

behind him, he is keen to exhibit his life's work to the public. He says he wants to both remind the West of its debt to the ancient Islamic world, and raise Arabs' pride in their past achievements.

Western science historians know very well that the Arab world was the guardian of the ancient Greeks' scientific knowledge during the Middle Ages, before the European Renaissance rediscovered and extended it. But they also acknowledge that there has been too little study of how the Arabs developed and used that knowledge, and how it fed Europe's cultural revival. "It has been a bit of a blind spot," says Peter Damerow of the Max Planck Institute for the History of Science in Berlin, "mostly because it demands such language skills."

In his youth, Sezgin trained under the sternest of taskmasters in his home town of Istanbul — the orientalist Hellmut Ritter, who insisted that his students learn a new

language every year. Thanks to such discipline, Sezgin is one of a small number of experts who has few linguistic problems with manuscripts written in Babylonian, Greek, Latin or Persian — in addition to Arabic.

Text messages

By the sixth century AD the scientific traditions of Ancient Greece had largely petered out, but the Greeks left behind important scientific and philosophical texts. For two centuries between 750 and 950, the caliphs of the Abbasid dynasty, centred in Baghdad, supported a mammoth translation effort. Greek works translated into Arabic include Euclid's geometry, Ptolemy's astronomy, the medical works of Galen and Hippocrates and the pharmacopoeia of Dioscorides.

The caliphs understood the importance of scholarship to their expanding empire. As the reach of the Islamic world spread, stretching from northern India to Spain, they absorbed

as much knowledge as they could from each conquest. In Persia and India, unlike Greece, the scientific traditions were still very much alive. And so it was to Persian and Indian scholars that the translators turned when they needed to make scientific, as well as linguistic, sense of the old Greek manuscripts.

Such expertise was vital because the caliphs wanted their acquired knowledge to deliver practical as well as intellectual benefits to their empire: from monumental architecture and city planning to medical care and transport. Luckily for Sezgin, this means that many texts written by Arab scholars include detailed engineering information on how to build mechanical devices, scientific instruments or architectural components.

In the fifteenth century, the Islamic world shrank under military pressure from western Europe — the last Muslim forces were forced out of Spain in 1492, the year Christopher Columbus reached America. By this time, the European Renaissance was under way and Islamic knowledge was sucked up by powers on the rise, such as Spain and France. Many Arabic works had by then been translated into Latin, but the sources themselves were neglected. Although European libraries and museums collected Arabic scripts, they sat in obscurity as they were largely indecipherable. Over time these trophies of ancient Islam were taken even farther afield to Russia and the United States.

Paper trail

The European Renaissance's debt to Islamic science has never been entirely forgotten, but there has been little systematic analysis. Sezgin is one of the few contemporary champions of the field. He spent more than three decades travelling the world seeking out 'lost' manuscripts, copying them and interpreting them. He tracked down some in countries formerly under Islamic rule, particularly India — which Sezgin estimates to have 50,000 such manuscripts.

His odyssey led to a 12-volume academic work, *Geschichte des arabischen Schrifttums* (*History of Arabic Literature*). Then, about 30 years ago, he decided to begin reconstructing the instruments he had read about. "Sadly, few had survived the centuries," says Sezgin.

His first reconstruction was a model of a horse-drawn ratcheted wheel to raise water from wells, built by the workshop at the University of Frankfurt, where Sezgin was appointed professor in 1966. It had been described in a book dating to about AD 1200 by Ibn al-Razzāz al-Jazari. As his enthusiasm deepened, Sezgin found himself casting far afield for craftsmen who could help reconstruct more complex instruments. The drawings of figures and pictures on a 1044 celestial globe were done in Bremen, but the

Jumbo timepiece

This 8-foot-high reconstruction of wood and brass keeps time perfectly, as Fuat Sezgin is happy to demonstrate. Inside a water-filled bucket, hidden within the elephant's body, floats a deep bowl with a small hole in its centre. The water leaks through the hole slowly, taking exactly 30 minutes to fill the bowl enough so that it sinks. As it does so, it pulls on a string attached to a 'see-saw' in the tower on the elephant's back, and the rest of the clock springs into action.

First, the see-saw releases a ball from the tower that falls into the mouth of one of the snakes. Under the weight, the snake tips forward and, through a system of strings, the figure in the tower raises his right or left hand to indicate the full or half hour, while another figure strikes his drum. Then, as the snakes' tails rise, the sunken bowl is hauled up to start the cycle all over again.



fine engraving, with the accompanying script, was done in Cairo.

Some of the items in his collection are fairly simple — the diverse surgical scalpels and cauterizing instruments, for example. But others, including astronomical instruments, are extraordinarily complex. Some of the most dramatic are enormous, intricate clocks that use water to measure time. Sezgin is particularly proud of an elaborate water clock described by al-Jazari in about 1200 (see 'Jumbo timepiece', above).

Sezgin is most animated when talking about ancient Arabic geographical and nautical scholars. But there are no true disciplinary borders to his pride. He beams just as broadly on recalling his discovery of manuscripts describing astronomical instruments used in tenth-century Islamic observatories — some six centuries before they were used in Europe.

And he gaily reassures visitors recoiling from his gruesome surgical instruments that "the old Arab world had general anaesthetics — morphine-based". He describes his collections of sophisticated Islamic weaponry, and the equally sophisticated distilleries for rose water, with ebullient pleasure. The museum also includes a major collection of original Islamic musical instruments, assembled by Sezgin's colleague Eckhard Neubauer.

Descriptions of the al-Jazari water clock were accompanied by detailed drawings. But other equipment was harder to reconstruct. For example, the advanced compass that a Portuguese sailor came across while sailing the Indian Ocean was reconstructed from a description by a sixteenth-century historian. "It was not necessarily much harder to reconstruct without pictures, but it always required a lot of reading of the text to be sure we had understood exactly," says Sezgin.

Whatever the challenge, Sezgin rose to it, sparing no time, trouble or money. He has

always used materials that would have been used at the time — heavy woods, brass, even gold when called for. When he retired, Sezgin created a foundation to raise money for the museum. But he has thrown his own money in too, saying that he lives as modestly as he can: "I don't have many personal needs." His total investment in the collection is well above €2 million (US\$2.7 million), he estimates, not including travel costs. How much might it be worth on the open market? "Maybe €50 million," says Sezgin, with a delighted laugh at the absurd thought that he could let it go at any price.

Hidden treasures

Despite its dimensions and diversity, the collection remains, even among scholars, almost unknown. Sezgin has chosen to be a loner in his venture. But this year, for the first time, he allowed a few small exhibitions to take place. Some of the instruments were shown at the Topkapi Palace Museum in Istanbul and at the Jewish Museum in Frankfurt. And there is now a virtual museum in German on the Internet.

Scholars who haven't seen the museum are intrigued. "Having physical instruments can be helpful in getting people to understand what they cannot read because of the language issue," says John Heilbron, a science historian at Worcester College, Oxford, UK. And even for those who can read the texts, says Damerow, "it's hard to know just how easy an instrument might have been to use, and how precise it may be in its measurements — it really helps to have it in your hand".

These scholars, and the wider public, may not have to wait much longer to see Sezgin's treasures up close. The first major exhibition of the collection is planned for spring 2006 at the Arab World Institute in Paris. Sezgin is finally opening his museum doors to the future, as well as to the past.

Alison Abbott is *Nature's* senior European correspondent.

► <http://web.uni-frankfurt.de/fb13/igaiw/museum/museum.html>

THE TOADS are coming!

Cane toads are infamous for wreaking havoc on Australian ecosystems. But, as Peter Aldhous discovers, we're only now about to learn whether their fearsome reputation is deserved.

Darwin is bracing itself for an invasion. *Bufo marinus*, otherwise known as the cane toad, has been advancing across the Australian landscape for decades. Ravenous, poisonous and an explosive breeder, it has acquired a reputation for wiping out this country's unique wildlife. Now the front line is just a few tens of kilometres from the capital of the Northern Territory; the first invaders are expected to hit the city's limits before the wet season ends next May.

This prospect horrifies Darwin's residents. They have visions of their pets being poisoned by the toads' potent toxins, while their swimming pools fill with the loathsome beasts. But as the cane toads approach Darwin, scientists are also being presented with their best opportunity to investigate the ecological impact of the amphibians. Despite the animals' infamy, this remains largely a matter of conjecture. "That's the whole thing with the toads," says Tony Griffiths, an ecologist at the Key Centre for Tropical Wildlife Management, based at the city's Charles Darwin University. "Most of it is anecdotal. There has been very little scientific evidence."

Until now, that is. With Griffiths and his colleagues David Bowman and Barry Brook, I'm surveying the front line from a helicopter, swooping over the huge flood plain of the Adelaide River, just outside Darwin. The landscape here is ideal territory for the toads, which hail from the wetlands of South America. "This is the closest thing to the Amazon flood plains that you are going to get in Australia," says Brook.

Cranes, geese and buffalo scatter beneath us, as we pick out the tracks of crocodiles in the mud. For decades, ecologists have flown over, walked through and scrutinized this plain and its neighbouring habitats. When the toads arrive — which they could do within weeks — they will be closely watched, providing the most detailed study yet of an invasion by a vertebrate on this scale. It's quite possible, says Bowman, that this will reveal that the toads' reputation for ecological vandalism has been exaggerated.

Cane toads were first introduced into



Under attack: Darwin (above) is about to be hit by a hoard of voracious cane toads (top left). Already, the northern quoll (left) seems to be falling victim to the amphibians' poisonous charms.

Queensland in 1935, to control beetles that were ravaging the sugar harvest. But the toads found other invertebrates more to their taste, and have since competed vigorously with native insectivores. They have also defeated most of their predators; the bulging poison glands behind their eyes mean that a first encounter with a cane toad is often the last.

But it's impossible to blame the toads definitively for the decline of any particular species. In the first few decades after the animals arrived, biologists were little concerned about the dangers they posed, and so didn't collect any data. More recently, the toads have been advancing to the northwest across

the remote Gulf of Carpentaria, where few ecologists venture. In fact, the toads attracted little scientific attention until 2001, when they showed up in Kakadu National Park.

This area of pristine wetlands and savannah forest, which enjoys World Heritage status, hosts some animals that are particularly sensitive to changes in the environment — including the northern quoll (*Dasyurus hallucatus*). "If ever there was a species highly adapted for extinction, it's the quoll," says Rod Kennett, Kakadu's project officer for ecology and natural-resource management. Seldom seen, the northern quoll is a rabbit-sized marsupial carnivore that will try to eat anything it





On the front line: (from left) Tony Griffiths, Barry Brook and David Bowman hope to learn more about cane toads' ecological effects as the animals spread across northern Australia (below).

There she is, bounding back and forth in the cage, large-eyed and inquisitive. Ansell manoeuvres her into a cloth bag, and she calms down. He identifies her from a microchip injected beneath her skin the first time she was trapped, and takes a series of measurements — weight, head length and so on. Her teats reveal that she is nursing about half a dozen young. Then she is released to scamper off into the night.

By the time we finish, around 2.30 a.m., we have caught two more females. The next night, we trap the same three animals — indicating that they represent the majority of the local population. The prognosis isn't good. This time last year, we might have captured 20 quolls each night, Ansell says. At Oakwood's other field site, to the southwest, the news is even grimmer. There, the quolls disappeared about 18 months ago. Several radio-collared animals were found dead with reddened gums — a tell-tale sign of cane-toad poisoning.

Kennett has little doubt that Kakadu's quolls are finished, so he is now concentrating the park's research funding on Griffiths's work on another vulnerable predator, the monitor lizard *Varanus panoptes*. These lizards have declined in number since the toads arrived in Kakadu. But there is at least some hope that populations might stabilize and bounce back — so Kennett wants to keep watch for future developments.

More signs of the toads' damaging effects come from a series of automated 'toad poles' in Kakadu — listening devices that pick up the calls of cane toads and native amphib-

ians. A team led by Gordon Grigg of the University of Queensland in Brisbane has preliminary data from these devices suggesting that the toads' arrival coincides with decline of native frogs.

This all sounds like damning evidence, but the big picture is likely to be more complicated. Even for the quolls, the toads may be just part of the story. Together with most of northern Australia's small mammals, quolls are in decline even in areas that haven't been invaded by cane toads. This is probably because of changes in fire management. For thousands of years, Australia's Aboriginal people burned the landscape frequently, but each fire was small, so any one patch would burn only once every few years. This created a rich mosaic of habitats, boosting biodiversity. Today's ranchers, in contrast, burn large areas each dry season.

Last suppers

These and other ecological pressures on the Darwin area (see 'On the wild frontier', overleaf) will all play a role in how the toads affect the ecosystems in and around the city. The extent to which an individual species is threatened by the toads will also depend on nuances of its ecology and behaviour. Some species of monitor lizard spend much of their time in water, for instance, which may limit their contact with the toads.

Researchers led by Rick Shine of the University of Sydney, a leading snake ecologist, are conducting some preliminary lab experiments to determine which of the species they study are likely to be most at risk. Of the non-venomous species, water pythons (*Liasis fuscus*) are probably safe, because they feed almost exclusively on rodents. Keelbacks (*Tropidonophis mairii*) seem reasonably resistant to the toad toxins, but slaty-grey snakes (*Stegonotus cucullatus*) are less so. The area's venomous snakes, meanwhile, have little resistance at all. "They are really going to be in trouble," predicts Greg Brown, a Canadian postdoc in Shine's team.

For such species, the hope is that they will cling on long enough for natural selection to lend a helping hand. Shine's graduate student Ben Phillips has examined museum specimens of snakes from Queensland, finding that vulnerable species may adapt over the generations to evolve smaller mouths and larger body sizes, making them less likely to eat a toad big enough to poison them (B. L. Phillips and R. Shine *Proc. Natl Acad. Sci. USA* **101**, 17150–17155; 2004). But the acid test will be what happens when the toads arrive at the team's main field site — an



can capture. This, combined with its unusual breeding habits, makes it especially vulnerable to the sudden appearance of an army of toxic creatures. The males all die each breeding season, exhausted by frenetic mating, and relatively few adult females make it from one year to the next, even in good times. Take out a small number of nursing mothers towards the end of the breeding season, and a local population can crash precipitously.

On the brink

So Kennett asked Meri Oakwood, who runs the consultancy Envirotek, based in Nana Glen, New South Wales, to keep an eye on the quolls. Unsurprisingly, she has found that they are not faring well.

Two nights spent at the northeastern boundary of Kakadu in mid-October brings this point home. Visiting the park with Shaun Ansell, a student at Charles Darwin University who works on Oakwood's project, I help to lay 50 wire-cage traps around a series of rocky outcrops, and bait them at sunset with a mixture of peanut butter, oats and vegetable oil — irresistible to small mammals. By 12.30 a.m., we've examined half the traps, finding nothing but rock wallabies and rodents. Then, Ansell shouts: "Quoll!"

artificial wetland created by a barrage known as Fogg Dam.

Countless nights walking the road along this dam have given Brown a supernatural eye for spotting snakes. I tag along with him for an evening's snake survey, and within a few minutes, his flashlight picks out a water python slithering through the undergrowth.

Brown later repeats the trick from the front seat of his four-wheel drive. As we negotiate the rough track across the nearby Harrison Dam, he spots one of each of his two main study species, the keelback and the slaty-grey. By the end of the evening, we've recorded a total of seven other water pythons. Not a bad night, says Brown. "But come back in a year, when the toads are everywhere, and who knows what you might or might not see."

Toad haul

Another project in the same area will examine the toads' appetite for invertebrates. One of Shine's undergraduate students, Matt Greenlees, has built a series of small metal enclosures in which he plans to put toads, to see how much damage they inflict on populations of insects and other invertebrates. Once the toads have colonized the Fogg Dam area, he will clear the enclosures and maintain them as toad-free habitats.

Out in the field at Fogg Dam with Brown, people driving by stop to ask what we're doing — and then want to know whether the toads can be stopped. The locals' concern is laudable, but has itself brought some adverse consequences for native wildlife. On a recent radio show, one caller gleefully described how she had driven back and forth over a mass of cane toads. But the toads had yet to hit her locality. She had probably squashed a bunch of superficially similar native marbled frogs (*Limnodynastes convexiusculus*).

So can anything be done? At the inaugural meeting of the National Cane Toad Task Force, held in Darwin in late October, there were few answers. Discussions about strategies for trapping toads to keep their numbers down in key habitats were plagued by the phrase "preliminary data". With so much in doubt, it's going to be difficult for the task force to devise a plan.

Those who have been calling for officials to address the issue for years can't conceal their fury. "Nobody lifted a bloody finger," complains Grahame Webb, who runs a wildlife research and education centre outside Darwin, and chairs the World Conservation Union's Crocodile Specialist Group.

Now a rearguard action is under way. The Northern Territory government, national-park officials and Aboriginal representatives are moving small populations of quolls to toad-free islands off the coast. If researchers find a way to wipe out the toads, the predators could then be reintroduced to the mainland.

But that's a distant prospect. Researchers with the CSIRO, Australia's national research

On the wild frontier

Today, Darwin is cosmopolitan, laid back, and a Mecca for backpackers. But not so long ago, it was a rough-and-ready frontier town — where people lived hard, drank hard and settled arguments with their fists.

Viewed from the air, it's clear that this is still a frontier town in another sense. Darwin is a small urban island in the midst of a largely untouched tropical wilderness. There are mangroves, flood plains and expanses of savannah forest that stretch as far as the eye can see. Patches of denser rainforest break up the scene, growing where groundwater lies closer to the surface.

So far, the city has done little to spoil this landscape. Indeed, its suburbs and satellite towns actually support a higher density of small mammals than the surrounding wilderness — probably because their spread-out development has created a patchwork of habitats (right) that mimics the beneficial pattern formerly created by ancient aboriginal burning practices.

But accelerating agricultural and industrial development could soon disturb this idyllic scene. Savannah forest is being cleared to grow cotton and other thirsty crops, which drain the groundwater. And a gas liquefaction plant is being built near the city's harbour. It is all part of the Australian government's policy of investing in its northernmost outpost.

Although the cane toad may be the best known threat to the habitats around Darwin, this burgeoning development could in the long run prove much more significant. "We're looking at an entire ecosystem being transformed under our feet," says David Bowman, who directs the Key Centre for Tropical Wildlife Management



P. ALDHOUS

at the city's Charles Darwin University. "In ten years, you won't be able to recognize the joint."

The good news is that, unlike other tropical landscapes facing similar problems, this wilderness is in a rich country with a habit of supporting ecologists keen on studying habitat change. Bowman and his colleagues are now applying for a five-year, A\$24-million (US\$19-million) grant to determine how various development pressures are affecting the region's landscapes and biodiversity. They hope to provide information that could allow Darwin's economy to grow in an environmentally sustainable way.

Whether that information will be used or not remains to be seen. First, says Bowman, developers and conservationists will have to step down from their diametrically opposed positions, and start listening to each other.



Snakes alive: Greg Brown holds a keelback, one of the species whose numbers are being counted in the face of the toad invasion.

agency, are trying to identify cane-toad-specific genes that control the development of tadpoles into toadlets. They then hope to engineer a ranavirus, which infects amphibians, to specifically disrupt these genes.

It might just be possible to tackle cane toads this way without threatening native amphibians: Australia has no other members of the family Bufonidae, to which the cane toad belongs. But there are bufonids in neighbouring southeast Asian countries. So any attempt to release such a virus could spark an international outcry — even if the Australian public was prepared to countenance the idea.

Bowman believes that the best hope lies with ecology, rather than genetic engineering. Northern Australia will have to learn to live with the toads, he argues. Ecological studies may reveal subtle ways of manipulating native habitats, through controlled burning perhaps, to minimize the damage they cause. Expecting a "cheap fix" of biological control to come to the rescue, Bowman suggests, is wishful thinking. After all, it was a misguided attempt at biological control that created the problem in the first place.

Peter Aldhous is Nature's chief news & features editor.

P. ALDHOUS

Genome news highlights loss of chicken strains

Future research is already under threat, as budget cuts wipe out irreplaceable lines.

Sir — Publication of the chicken genome sequence and comparative analysis last week (*Nature* **432**, 695–716 and 717–722; 2004), together with a map of more than 2.8 million single nucleotide polymorphisms, is a cause for celebration. The best way to make sense of these data is to manipulate the genes of living organisms. But worldwide stocks of the animals needed to pursue such gene-function studies are dwindling, and in the United States they are being lost to budget cuts. So the future of avian research looks dim despite the opportunities — for understanding the genetic bases for development, immunity, host–pathogen interactions and disease resistance — offered by the chicken genome project.

Avian stocks are being eliminated in university, government and commercial facilities faced with budget constraints. The crisis described in 2003 by J. E. Fulton and M. E. Delany (*Science* **300**, 1667–1668; 2003) has not been addressed, and additional stocks are being eliminated or threatened. Five different US sites have eliminated stocks in the past year alone. We estimate that at least two-thirds of the developmental mutant lines and more than half the inbred

lines held in North America are located in threatened facilities. Unfortunately, given the struggle to maintain living stocks, efforts towards developing new chicken lines for research has virtually stopped. This is in striking contrast to the situation in the mouse, where the generation of new genetic strains has become a priority at almost every major biomedical institution (see “Geneticists prepare for deluge of mutant mice”, *Nature* **432**, 541; 2004).

Often the decision to eliminate poultry lines is made with insufficient time to respond. Further, many genetic strains are now held at only one facility, making these vulnerable to loss from disease outbreaks such as avian influenza and exotic Newcastle disease. Today, the only US government support for conservation exists in the National Animal Germplasm Program (NAGP) of the US Department of Agriculture’s Agricultural Research Service. This programme, with \$679,000 funds per year, has a mandate to preserve species from cattle to aquatic organisms, mostly by cryopreservation. The NAGP currently has 162,000 units of animal germplasm; fewer than 1,300 (<1%) are chicken semen samples.

No avian embryo samples or ova are cryopreserved at the NAGP. The nature of avian ova means that current cryopreservation techniques are of limited value.

Both immediate action to preserve the remaining poultry resources and long-term, sustainable solutions are essential. An internationally supported plan needs to be developed and implemented for maintaining avian genetic stocks and exploiting them for both biomedical and agricultural research. Funding is needed to develop facilities — similar to those of the Jackson Laboratory in Bar Harbor, Maine, for mouse strains — to conserve important chicken genetic stocks, to develop means of ova cryopreservation, and to make stocks available for research. Replacing lost stocks, especially of rare inbred strains, will take many years at substantially greater cost than that of preserving what we have now.

Marcia M. Miller

*City of Hope National Medical Center,
Duarte, California 91010, USA*

*Signed on behalf of 46 international co-authors,
whose names and contact details are available at
<http://animalscience.ucdavis.edu/AvianResources>
and <http://poultry.mph.msu.edu>.*

Media affect opinions less than they would like

Sir — Your Editorial “Going public” (*Nature* **431**, 883; 2004) says “British scientists have seen the public swayed by misleading media coverage of genetically modified (GM) food and vaccines”. What evidence do you have for suggesting that there was more media bias one way or the other?

The evidence in the scientific literature suggests that public attitudes to GM foods in a number of European countries were formed independently of the scale and nature of media coverage (see S. Mayer and A. Stirling *EMBO Rep.* **5**, 1021–1024; 2004, and C. Marris, B. Wynne, P. Simmons and S. Weldon *Public Perceptions of Biotechnologies in Europe* at www.lancs.ac.uk/depts/ieppp/pabe/docs.htm).

This message will be as unwelcome to some on the anti-GM side (who fall into the trap of judging their success by the amount of coverage they get) as to pro-GM scientists and companies, who would rather think that their products were rejected because of unfair media coverage than simply because the public did not want them. Of course, the group to whom this conclusion is least palatable is journalists, who like to think their readers believe every word they write.

However, this ability to make decisions independently of media coverage supports your contention that scientists should trust the public to make a sensible contribution to discussions on research priorities.

Peter Melchett

*Soil Association, Bristol House,
40–56 Victoria Street, Bristol BS1 6BY, UK*

Brazil’s efforts to reverse decline in scholarships

Sir — I would like to update the details of Brazilian scholarships discussed in Correspondence by A. F. Helene and V. S. Valentinuzzi (*Nature* **431**, 627; 2004).

The downward trend in the number of MSc and PhD scholarships is being reversed by President Luiz da Silva’s government, inaugurated in January 2003. The number of PhD and MSc scholarships are: 5,917 and 5,881 (2001); 5,668 and 5,510 (2002); 5,947 and 5,985 (2003); 6,316 and 6,769 (2004).

In addition, in 2004 the National Council for Scientific and Technological Development (CNPq) has allocated an 18% increase to the scholarships, as an initial effort towards restoring their value.

Moreover, CNPq has added a monthly ‘bench grant’ of 30% on top of the

scholarships, to help labs buy specific items needed for students’ PhD work.

Erney Plessmann Camargo

CNPq, SEPN 507, Brasilia DF, 70740-901, Brazil

EPA not responsible for ‘uninvitation’ of a critic

Sir — Contrary to the claim reported in your News story, “Herbicide critic dropped from pollution conference” (*Nature* **432**, 136; 2004), an official of the US Environmental Protection Agency (EPA) did not “instigate” the “uninvitation” of Dr Tyrone Hayes in the conference sponsored by the Minnesota Pollution Control Agency. The EPA had no involvement in this decision. When Minnesota state officials contacted the EPA about this conference, the EPA only provided them with web links where interested parties could read about the EPA’s exhaustive evaluation of atrazine.

James J. Jones

Office of Pesticide Programs, US Environmental Protection Agency, 1200 Pennsylvania Avenue NW, Washington DC 20460, USA

The EPA was asked to comment before the News story was printed, but declined to do so — Editor, Correspondence.

Harnessing nature in Africa

Biological pest control can benefit the pocket, health and the environment.

Peter Neuenschwander

In Africa, swarms of desert locusts are threatening crops from the Atlantic to the Red Sea, and 12 million hectares of often fragile environments have been treated with insecticides. This is worrying, as the last major swarms in 1988 took several years, 1.5 million litres of chemical pesticides and US\$300 million to bring under control. Since then, ecologists, economists and politicians concerned about the effects of these pesticides on human health and the environment have agreed that alternative methods were needed to combat locust outbreaks — ideally before the locusts invade agricultural land in the Sahel region.

Such methods would replace chemicals with biological alternatives that are environmentally friendly and economically viable. After 15 years of work by an international collaboration (which includes my institute), a new biopesticide based on fungal spores, Green Muscle, has been licensed for use against grasshoppers and locusts in many African countries¹. It has been successful in field trials against grasshoppers in West Africa and is well received by farmers, but has not yet been widely deployed. Part of the problem is the relatively slow action and perceived high cost of Green Muscle. But the biggest hurdles are regulatory and the perception that biological control — whether in a crisis or for everyday use — is not a true alternative to synthetic pesticides.

The 'green revolution' of the 1960s and 1970s introduced high-yielding crop varieties, which flourished when combined with synthetic pesticides and high doses of fertilizer. This resulted in an increase in agricultural production in Asia and South America, but Africa did not benefit to the same degree. Since then, large-scale programmes for breeding pest-resistant, high-yielding 'improved' varieties that do not depend on high quantities of fertilizer or insecticide have brought considerable benefits to Africa², with estimated productivity gains of US\$335 million for cassava (West and Central Africa alone) and US\$1.33 billion for maize. Improved plant resistance remains the most important tool in the fight against viruses, bacteria, fungi and nematodes, and is one of the four pillars of integrated pest management (IPM). The other three are biological, chemical and cultural (such as ploughing, weeding and mulching) control.

Improved crop varieties can provide a complete solution, at least temporarily, when used in conjunction with good soil



Looking for solutions: biopesticides could prevent locust swarms.

management. By keeping pest populations at a minimum, they reduce the likelihood that pests will overcome plant resistance mechanisms and prevent other potential pests from becoming a serious threat. But developing and then getting producers to use improved varieties and new farming practices is a slow and painstaking exercise.

New pesticides, when integrated into a sound IPM system, are less harmful than old products, but most African farmers cannot afford them. Instead they use cheaper (sometimes obsolete) chemicals — some of them banned in Europe and North America. Consequently, misuse of synthetic chemicals is rampant, causing well documented health and pollution problems, particularly on cowpea and stored maize.

Battle of the beasts

It is time to consider other options that offer economic and environmental benefits. With foreign species invading Africa at an increasing rate and threatening agriculture and conservation, we argue that biological control, particularly against insects, mites and weeds, should receive more attention. Classical biological control consists of the introduction, establishment and spread of specific natural enemies against exotic pests. Proponents of this approach are sometimes accused of being 'green dreamers', who want to keep poor farmers poor. But the data in Table 1 demonstrate that even if you consider the impact of these biological control technologies in terms of their yield only — without taking into account the environmental impact — they offer benefits comparable to those of the long-term plant-breeding programmes for cassava and maize². A closer look at the strengths, weaknesses and ecological context of biological control is needed.

Before they invaded Africa, some exotic pests, such as the two mealybugs, were not even known in their native countries. By contrast, water hyacinth and other water weeds, which impede fishing and traffic (including transport of agricultural produce), are also pests in their homelands. The benefit:cost ratios for these biological control projects are astonishingly high, far above 100. The successes

in Table 1 have been evaluated using modern economic models, taking into account depreciation (6–10%, depending on the project). Cassava green mite, mango mealybug and water hyacinth are now controlled with minimal additional costs in many other countries. Many more examples of successful biological control (but without corresponding economic evaluation) have been documented across Africa, for practically all crops³.

This type of pest control has a unique advantage: it is self-regulating. Unlike resistant crop varieties, biological control pits pests against their natural enemies. In a continuous arms race, these parasitoids and predators counteract any emerging defence mechanisms, thus assuring that biological control does not break down. When 'resurgences' of previously well controlled insects occur, these are most often attributable to synthetic insecticides applied against other targets, which have killed exotic and indigenous natural enemies indiscriminately. In the past, some biological control agents had unfortunate effects on non-target populations, but today such problems are rare⁴, despite claims to the contrary. Because biological control agents are screened in quarantine laboratories before their release, they are free of their natural enemies (particularly diseases and so-called hyperparasitoids) so their impact can be greater than at their place of origin. Success is also affected by interactions between biological control agents and crop varieties, as seen with an effective green mite predator that finds refuge in the hairy tips of some cassava varieties.

Adapting to Africa

Unfortunately, farmers and politicians alike easily forget the impact of biological control once arthropod and weed populations have reached acceptable levels. In addition,

Table 1. Economic impact analysis of ongoing biological control projects in Africa: the estimated savings were achieved for costs far below 1%.

Pest species and year of first occurrence	Typical losses in yield	Biological control agent	Start of campaign	Area under economic analysis	Reduction in loss	Estimated savings in US\$ million
Cassava mealybug 1973 (ref. 7)	40%	Encyrtid wasp <i>Anagyrus lopezi</i>	1981	27 African nations	90–95%	7,971–20,226
Cassava green mite 1971 II	35%	Phytoseiid mite <i>Typhlodromalus aripo</i>	1983	Nigeria, Ghana, Benin	80–95%	2,157
Mango mealybug 1980s (ref. 8)	90%	Encyrtid wasp <i>Gyransoidea tebygi</i>	1987	Benin	90%	531
Water hyacinth 1980 (ref. 9)	66% *	Weevil <i>Neochetina eichhorniae</i>	1991	Benin	36% †	260
Red waterfern 1978 (ref. 10)	‡	Weevil <i>Stenopelmus rufinasus</i>	1997	Republic of South Africa	§	206

* Damages of US\$84 million to fishing and trade at peak of infestation. † By 1999, full impact not yet achieved. ‡ Average damages of US\$533 per respondent (30 in total). § After three years the weed was not considered a problem anymore. II O. Coulibaly and R. Hannah, personal communication.

achieving pest control — finding equilibrium levels — often takes many years and the result is not the instant ‘extermination’ that is wished for by many farmers. But although not all pests are amenable to classical biological control, many more species should be targeted. Certain pan-tropical pests — which had invaded the tropics long before entomologists first described them — require world-wide exploration to find their native country and potential natural enemies. Cereal-stem borers are good examples of pests against which regional exchange of parasitoids has proved particularly promising⁵.

Although the mass release of beneficial insects is sometimes profitable in Europe and North America, this approach has generally failed in tropical Africa because of the high cost of insect-rearing facilities. This situation is unlikely to change in the near future. The mass release of insect pathogens, such as Green Muscle, shows more promise because of the greater ease of production and storage of the fungal spores. Finding an effective pathogen is one thing, but there are many additional challenges, such as creating production capacity, identifying key markets, solving conflicts over intellectual-property rights, and harmonizing regulatory and quality-control procedures. None of these were easy with Green Muscle.

Although the Food and Agriculture Organization of the United Nations ranked Green Muscle top in both environmental compatibility and lack of negative impact on human health, it still has not been widely used in Africa beyond the countries participating in its development, mostly Niger and Mali, where it has been deployed against Sahelian grasshopper outbreaks since 2000. On a smaller scale, it was used in field trials against grasshoppers in Benin, Cabo Verde and some east African countries. There have been fewer field trials against locusts. The absence of a uniform regulatory framework in a fragmented market is largely to blame for the limited sales. Existing regulations concern synthetic pesticides and are slowly being adapted to biopesticides. It also does not help that so many countries are involved, each with its own authority in such matters, and that the continental agency for coordinating plant protection legislation and

interventions, the Inter-African Phytosanitary Council, has only an advisory capacity and needs more support.

Proof that technical problems do not limit the use of Green Muscle is shown by the successful continent-wide application of an Australian version of Green muscle, Green Guard. The environmental benefits of Green Muscle, including the protection of birds migrating through the Sahara, should appeal to donors who previously provided other insecticides for free. Therefore, the question is whether concerned European and African governments and international donors can muster the political will to implement this new solution, which they asked for back in 1989.

Natural advantages

The same argument could apply to other environmentally friendly entomopathogens (insect viruses, bacteria or fungi). Although agents that target a narrow spectrum of species with satisfactory killing power are common, this is not sufficient for their successful deployment in the field. As seen with Green Muscle, the creation of production capacity and other problems need to be solved. Farmers interested in growing vegetables for export find the use of entomopathogens particularly attractive, as European Union regulations impose near zero-tolerance limits on pesticide residues.

Another promising approach to biologically oriented pest control involves insecticidal plant extracts, known as botanicals, which are easily available. Local African knowledge by botanists and elders is disappearing and often does not reach the farmer. Although some of these plant products are as poisonous as synthetic insecticides and would need regulatory testing, others, such as neem seed or leaf extracts, have demonstrated no negative side-effects⁶. But use of these extracts is still not widespread — despite the abundance of their source plants — because their commercial production by small-scale entrepreneurs has not yet taken off, as it has in India⁶.

These three options, classical biological control agents, entomopathogens and plant extracts (if screened correctly), share several advantages: they are fully compatible with new

crop varieties and cultural practices, and they preserve indigenous natural enemies. Unlike the extensive monocultures of the developed world, the small African fields, rich in other plants (‘weeds’) and often surrounded by strips of native vegetation, provide a welcoming environment to biological control agents.

In Africa, increases in production for the past 40 years have barely kept pace with population growth. They were mostly obtained by increasing cultivated areas through expansion into marginal, often biodiversity-rich and sometimes protected areas. As fallow periods are being phased out across Africa, with negative effects on soil regeneration and biodiversity, production on existing cultivated land must be intensified by using high-yielding varieties, a careful mix of organic and inorganic fertilizer, improved pesticides, and by maintaining natural enemies. Without help for subsistence farmers to buy fertilizers, it is likely that they will be pushed further into marginal or protected areas.

The eco-friendly options discussed here offer sustainable yield increases and allow us to get away from unsustainable insecticide use. Implementation of such programmes requires better education of stakeholders, best achieved by the many types of IPM farmers’ field schools. In addition, international collaboration, public investment under fair-trade practices, political will and public support need to improve. With a clear vision, a pragmatic approach and good leadership, many countries could adopt this approach today. ■

Peter Neuenschwander is at the International Institute of Tropical Agriculture, Biological Control Center for Africa, 08 B.P. 0932 Cotonou, Benin.

1. Lomer, C. et al. *Annu. Rev. Entomol.* **46**, 667–702 (2001).
2. *Crop Improvement and its Effect on Productivity: The Impact of International Agricultural Research* (eds Evenson, R. E. & Gollin, D.) Ch. 8 & 16 (CAB International, Wallingford, 2003).
3. Neuenschwander, P., Borgemeister, C. & Langewald, J. (eds) *Biological Control in IPM Systems in Africa* (CAB International, Wallingford, 2003).
4. Lynch, L. D. & Thomas, M. B. *Biocontrol News and Information* **21**, 117N–130N (2000).
5. Kfir, R. et al. *Annu. Rev. Entomol.* **47**, 701–731 (2002).
6. Schmutterer, H. (ed.) *The Neem Tree* (Neem Foundation, Mumbai, 2002).
7. Zeddes, J. et al. *Agric. Econ.* **24**, 209–219 (2001).
8. Bokkon-Ganta, A. H., de Groot, H. & Neuenschwander, P. *Agric. Ecosyst. Environ.* **93**, 367–378 (2002).
9. de Groot, H. et al. *Ecol. Econ.* **45**, 105–117 (2003).
10. McConnachie, A. et al. *Biol. Control* **28**, 25–32 (2003).



CHRISTIAN DARKIN

Echo of the big bang

An end to the boom in popular science books may actually raise standards.

Peter Tallack

Stephen Hawking may have taken science to the bedside table, but the phenomenon is over. The attitude of publishers and book-sellers towards popular science writing has changed, and the market is not as buoyant as it was a decade ago. Gone are the days when publishers were falling over themselves to acquire the next *Brief History of Time* or *Longitude*, and were happy to pay big bucks to fend off the competition.

Whereas authors could once command six-figure advances for books on cutting-edge science or narrative-driven histories of the subject, most today struggle to get a deal, let alone shift several thousand copies of their books. This is not to say that the audience for serious popular science has evaporated: there are still plenty of readers hungry for authoritative, well-written, interesting books. But the 'trade' — the commercial sector of the publishing market — has raised the bar for what kinds of book find their way on to the bookstands.

Why has this big bang turned out to be a whimper? Part of the problem is that the trade has decided that popular science as a genre is not what it was cracked up to be. So many people — academics and journalists, as well as publishers and agents — jumped on the bandwagon that the market became saturated and publishers got their fingers

burnt. The public is spoilt for choice when it comes to books on genetics or cosmology, mathematics or neuroscience. Just how many books do they really need on the impending threat from asteroids and comets, the sequencing of the human genome, or the challenge of the Riemann hypothesis?

The pace of popular science publishing has outstripped that of scientific advance — even popular-science aficionados would be hard pressed to name one scientific leap in our understanding of human consciousness, despite the vast number of books on the subject. Big ideas simply don't come along very often, and when they do, several books on the same idea often get signed up simultaneously. Witness for instance the recent spate of titles on network theory by Mark

Buchanan, Duncan Watts, Albert-Laszlo Barabasi, Philip Ball and Steven Strogatz. Despite the impeccable credentials of the authors, none of these has taken off, because there is a limit to how many similar titles the market can sustain.

This means that publishers today must be absolutely sure that every science book they commission has what it takes to stand out from the crowd. They are placing ever greater store on works that are original, topical and important — and written with verve and style by authors who are in full command of their subjects. And with publishing success nowadays so dependent on marketing and publicity, it is an advantage if authors have some kind of public profile. That is how Bill Bryson's funny but otherwise pedestrian overview of science, *A Short History of Nearly Everything*, managed to reach the parts that many established science writers couldn't reach — the bestseller lists.

But as well as a lack of public demand, there is a more pernicious force conspiring to keep popular science at bay: the handful of buyers in the head offices of the key book-shop chains, who decide which titles their stores should stock. This is in stark contrast to a few years ago, when most buying was done by the branch staff, and has altered the complexion of publishing as a whole.

A publisher at a major UK trade house, for example, was recently told that the 'key

Scientist wins Guardian First Book Award

Armand Marie Leroi's book *Mutants: On the Form, Varieties and Errors of the Body* has won the Guardian First Book Award 2004. This prize rewards new writing across fiction and non-fiction, and past winners include Zadie Smith for her novel *White Teeth*. Peter Little, reviewing Leroi's book for *Nature* (427, 101–102; 2004), wrote: "*Mutants* is an exquisitely life-enhancing book. It captures what we know of the development of what makes us human... Read it and enjoy words written carefully, elegantly and with sensibility."

accounts' had decided that scientific micro-histories were 'over', and he was thus finding it increasingly difficult to take anything on at all in this genre. This is despite a lack of real evidence that the public's appetite for books such as Simon Singh's *Fermat's Last Theorem* or Mark Kurlansky's *Cod* is satiated.

The upside to all this is that there has been a move back towards traditional history — epics such as Deborah Cadbury's *The Dinosaur Hunters* or Jenny Uglow's *The Lunar Men*. This trend is hardly surprising. In an increasingly weighty world, there is an increasing demand for weighty books, including agenda-driven titles such as Eric Schlosser's *Fast Food Nation*, Bjørn Lomborg's *The Skeptical Environmentalist*, and even aspirational titles, such as Roger Penrose's *The Road to Reality*, which offers nothing less than a complete advanced course in modern physics.

The market for popular science is still there — but as an echo of the original big bang. Publishers are increasingly sophisticated and discerning, and there has been a shakedown in the number of commercial houses who know, like and succeed with science. Ironically for an agent, I happen to think this is good news. It means that authors are taken on only by genuinely committed editors, and this in turn means that when their books do appear, people are more likely to buy, read and talk about them — more a case of Hawking radiation than the Hawking phenomenon. ■

Peter Tallack was formerly the book-review editor of *Nature and science* publishing director of Weidenfeld & Nicolson. He is now a partner of the literary agency Conville & Walsh, 2 Ganton Street, London W1F 7QL, UK.

The search for meaning

The Dynamic Dance: Nonvocal Communication in African Great Apes

by Barbara J. King

Harvard University Press: 2004. 240 pp.
\$29.95, £19.95, €27.70

W. Tecumseh Fitch

The idea that certain primate vocalizations are information-bearing signals is an old one. In 1892, R. L. Garner used playback of monkey vocalizations to deduce that some monkey calls had referential significance — they were what we now know as food or alarm calls. But it was much later, only about fifty years ago, that the concepts of 'information' and 'signal' became clear enough to be formalized mathematically by Norbert Wiener and Claude Shannon, giving birth to modern information theory.



The kiss: chimpanzee communication is based on interactions, according to dynamic systems theory. (Photographs courtesy of Frans de Waal.)

By isolating and formally defining a quantity termed information, which has surprising affinities to the physicists' concept of entropy, Shannon and Wiener planted the seeds of today's digital world, where diverse types of information can be transformed, stored or transmitted as a pattern of binary digits (or 'bits', a term they introduced). Shannon and Wiener were acutely aware that information (a measurable quantity of signals) is not to be confused with meaning (which depends on context and interpretation, and exists in the eye of the beholder). They both explicitly set aside 'meaning' as a topic for future work, and it remains formally undefined today.

Because meaning, rather than information, is central to animal communication, information theory plays a less prominent role in contemporary ethology than in telecommunications research or neuroscience. Indeed, there is a growing tendency among some primatologists to reject information theory entirely, and Barbara King's book *The Dynamic Dance* is a forceful example of this trend.

King is an anthropologist who has spent

years observing captive chimpanzees, bonobos and gorillas. The first 85 pages of this book argue that the 'signaller–receiver' model of animal communication (King calls it the Shannon/Marler framework, acknowledging the influence of ethologist Peter Marler) must be supplanted by King's new approach, dynamic systems theory. Dynamic-systems theorists eschew quantitative data, reject the notion that animal signals contain information, and focus instead on interactive aspects of ape social behaviour (their 'dynamic dance').

The core insight in King's book is that communicative exchanges among apes are 'co-regulated': interactions unfold contingently and unpredictably. King is certainly correct that ape communication systems (like those of many other animals) are complex and contingent. Far from being simple robotic automata that respond to a particular signal in a fixed and pre-programmed fashion, apes show considerable social intelligence, interpreting each other's actions against a backdrop of their history and current social context, and responding appropriately. But this is also true of two dogs interacting, as is quickly apparent to even casual observers.

Recognition of contingency and context in communication provides a rationale not for rejecting information theory, but for extending it, as its founders recognized. Information theory rigorously specifies a quantitative upper bound to the information borne in signals versus that supplied by receivers — surely a useful tool for understanding how receivers interpret and respond to communicative behaviour. I have rarely seen the baby–bathwater distinction so consistently elided as in this book.

The second part of the book features detailed descriptions of ape behavioural interactions, many based on King's own observations. King studiously avoids any interpretations of ape intentionality. Furthermore, because she rejects quantitative data (which she believes obscure the underlying co-regulatory nature of communication), we are given no summaries or statistics. Unfortunately, these two theoretical convictions conspire to make this section rather heavy going.

Space precludes quoting these descriptions in full, but here is a short excerpt of four from a sequence of nineteen behaviours: "Elikya watches Tamuli eat, then pulls on Tamuli's hand... As Tamuli chews, Elikya attempts to take a bit of orange but Tamuli turns her head away. Elikya again tries to take a bit of Tamuli's chewed orange but Tamuli pulls her head back; Elikya climbs off Tamuli and goes back to her mother." The middle 100 pages of the book are full of such "qualitative data". Although King is undoubtedly a sensitive observer of ape interactions, she rarely shares her

interpretations of these sequences with the reader, and these descriptions are presented primarily as demonstrations of the co-regulated nature of ape interactions. Although the cover describes the book as “eye opening”, I personally found these long and detailed descriptions of ape interactions made excellent bedtime reading.

In conclusion, like most clarion calls to scientific revolution, this book both oversimplifies the opposing viewpoint and overstates the significance and novelty of its own stance. King is correct that information theory should be extended to incorporate context and meaning, but this point was noted by the theory’s founders and is widely recognized today. The reader seeking an alternative in ‘dynamic systems theory’, as presented in this book, will find little more than a promissory note, and would probably be better served by spending a day at the zoo watching ape social behaviour themselves. ■

W. Tecumseh Fitch is at the School of Psychology, University of St Andrews, St Andrews, Fife KY16 9JU, UK.

Back to the roots

Ethnoflora of the Soqatra Archipelago

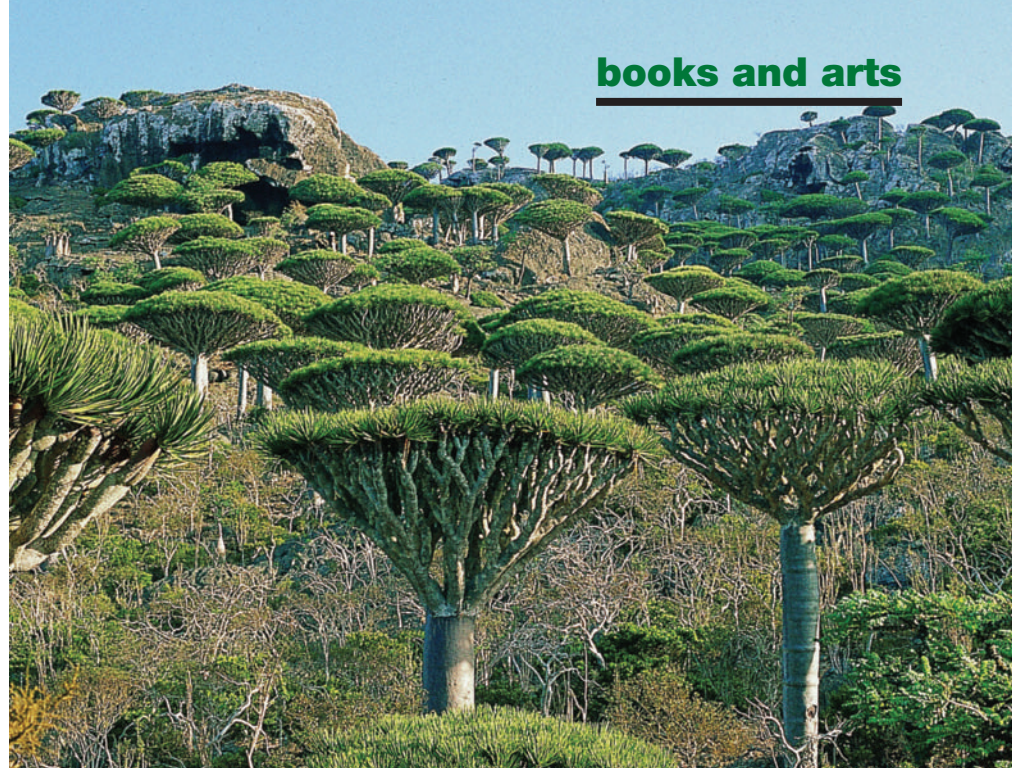
by Anthony G. Miller & Miranda Morris
Royal Botanic Garden Edinburgh: 2004.
759 pp. £75

Jan Salick

Stunning and sumptuous in design, *Ethnoflora of the Soqatra Archipelago* sets a contemporary standard for describing the plant life of an area and its uses. And what better focus than the exotic, unique and rich flora of the Soqatra Archipelago, south of the Arabian peninsula, off the eastern horn of Africa? An ethnoflora, although described in the book simply as the product of botanists working closely in the field with ethnographers and linguists, surpasses this process by matching diverse disciplines, objectives and issues. It combines traditional and scientific knowledge; botanical and ethnographic–linguistic description; conservation and sustainable management; and geography, plants and people.

Of the people, and for the people, of Soqatra, this work is accomplished here by an eminently qualified team from the Royal Botanic Garden Edinburgh. Anthony G. Miller is the foremost botanist of the Arabian peninsula; Miranda Morris is a linguist and ethnographer of the southern Arabian peninsula; Diccon Alexander is an illustrator, designer and photographer; and Ruth Atkinson is an ecologist, taxonomist and editor of this volume.

This ethnoflora differs from traditional



Local knowledge: ‘dragon’s blood’, the resin from *Dracaena cinnabari* trees, is used to fight infections.

floras in several ways. Comprehensive ethnographic information is presented for all species used by the islanders and their livestock. The fully illustrated key, figures and photographs illustrate all of the species and many of their characters — even the most basic, such as leaf position. Conventional taxonomic descriptions are brief and include minimal botanical terminology. These characteristics make the book meaningful and useful to a wider audience than professional botanists, including local people (if they can read English). For the benefit of conservationists and sustainable developers, all of the endemic species in the book are rated using the World Conservation Union’s threat categories.

Soqatra is home to an ancient, useful flora and 50,000 traditional people who have sustainably managed their unique centre of biodiversity and their environment for millennia. So at a deeper level, this emphasis on traditional sustainable use and management is a theme that fundamentally distinguishes this work from other floras.

The Soqotran flora cannot be discussed without mentioning the astounding plants and the landscape in which they thrive. ‘Dragon’s blood’ resin, for example, is collected from the plant *Dracaena cinnabari* and is used to fight infection and inflammation. Bitter aloes — the juice of *Aloe pernyi* — has many important pharmaceutical and medicinal uses. The latex of the Soqotran desert rose (*Adenium obesum sokotranum*) is often applied to wounds. And how weird is the cucumber tree, *Dendrosicyos socotrana*? On Soqatra, these bizarre plants exist in an almost surreal environment. The endemic flora of Soqatra — one-third of its entire flora — either split from Gondwana when the ancient southern supercontinent fragmented,

or migrated to the islands from neighbouring continents and radiated to form a unique and fantastic flora.

This weighty book cannot be reviewed without extolling its stunning beauty. There is a plethora of coloured photographs, original line drawings, illustrated keys, maps, images and historic drawings. The pages are handsomely designed and never crowded. The text is easy on the eye (although the location maps are somewhat faint) and the book’s organization is readily apparent and consistent.

It is difficult to find weaknesses with this exceptional book. However, my taxonomic colleagues are always up to the challenge: they have pointed out some problems with the one big key they tried (group G) and question the accuracy of some sketches in the keys. They also challenge some of the details about certain families: for example, recent molecular studies show that *Sisymbrium* includes 45 species, not 90. Of more concern to me, as an ethnobotanist attempting to turn a historically descriptive field into a modern science, is the predominance of description and lack of analysis in the book.

In a US National Science Foundation workshop, “Intellectual Imperatives in Ethnobiology”, held in the Missouri Botanical Garden in 2002, we emphasized the need “to explore modern methodology appropriate for studying people–biota–environment interactions; to quantitatively analyze our multidisciplinary data based on hypotheses; [and] to develop interdisciplinary education programs to train students and practitioners of ethnobiology.” In *Ethnoflora of the Soqatra Peninsular* there is little evidence of modern methods, analyses or even education extended to Soqotrans. Additionally, many ethical issues — intellectual–property rights,

a recognition of intellectual contribution, equitable distribution of benefits, capacity building and so forth — are not generally considered in this book.

This ethnoflora is set in the purely descriptive and colonial ethnobotanical traditions dating back to Linnaeus and before. Nonetheless, it is most useful for documentation, conservation and development, even if it does not advance the modern science of ethnobotany or reconcile inequities between scientific and traditional knowledge. ■

Jan Salick is curator of ethnobotany, Missouri Botanical Garden, Box 299, St Louis, Missouri 63166-0299, USA.

Reforming research

Scrutinising Science: The Changing UK Government of Science

by Rebecca Boden, Deborah Cox, Maria Nedeva & Katherine Barker

Palgrave: 2004. 218 pp. £45

Brian Heap

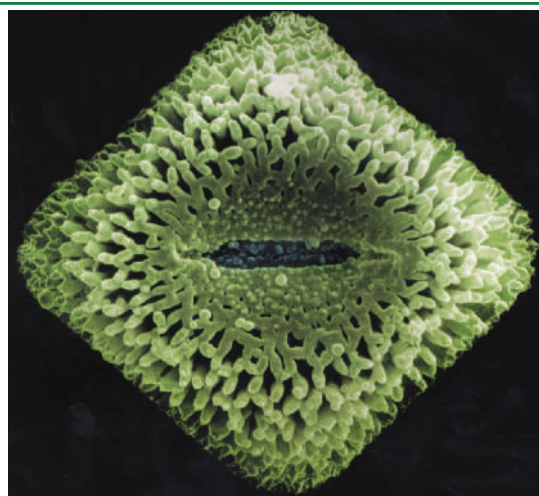
The dramatic changes in UK government research establishments (GREs) during the past 20 years constitute one of the most radical experiments in the organization and management of scientific research. And while these changes have been going on, mad-cow disease, foot-and-mouth disease and anthrax have thrust these labs into the public view. At other times, however, their work is less conspicuous but nonetheless highly significant for the public good.

During the 1980s and 1990s, successive Conservative governments seized the opportunity to address the perceived shortcomings of GRE laboratories. The experiment is outlined in the book *Scrutinising Science*, where it is seen through the lens of social and political science. The authors' analysis is based on interactions with policy-makers, senior managers and administrators tasked with carrying out their paymasters' orders, rather than on the firsthand experience of bench scientists. Some scientists found that the trauma of uncertainty and indecision resulted in planning blight and a loss of morale that was not readily reversed.

The authors helpfully trace the historical origins of the reform of GREs to theories that viewed science as a social institution: governments and politicians were supposed to fund science generously but should not intervene in the affairs of science, nor expect any tangible returns. The research councils that were formed early in the twentieth century existed on the principle of 'research council autonomy' espoused by R. B. Haldane, and other research establishments were created to serve their government departments or ministries directly. In the second half of the twentieth century,

Secret sex

Invisible to the naked eye, pollen grains are tough enough to have survived for millennia and display a wide diversity of forms, fascinating artists and scientists alike. In *Pollen: The Hidden Sexuality of Flowers* (Papadakis, £35), botanist Madeline Harley and artist Rob Kesseler bring pollen into closer view. The informative text is accompanied by stunning original micrographs of pollen grains, and photographs of the flowers that shed them. The book also includes engravings by previous observers of pollen, including Franz Bauer and Ernst Haeckel.



new models of governance and ownership structures appeared, or were imposed upon GREs. These models ensured closer control by departments through measures such as contractor–customer arrangements — part and parcel of making science and technology 'useful' to the departmental customer.

Later, an emerging 'new public management' (NPM) ideology pervaded government policies, driven by the practice of management by accounting. The NPM reform was driven with "breathless urgency" by politicians such as Michael Heseltine, with an outcome typified by complexity and heterogeneity. Eighteen science and technology establishments were transformed into 'next step agencies' between 1989 and 1996, with funding allocated by the contractor–customer mechanisms; 15 establishments were translated into 'executive agencies' by 1992. The process was suspended in 1997 by the incoming Labour government. The authors conclude that the application of NPM reform to GREs was complicated, messy and driven by opportunism, because GREs were a relatively minor part of government.

What of the present? This complicated picture is addressed by a brief description of individual GREs and their current status. There is also a discussion of eight laboratories that have undergone this transformational organization, including the National Engineering Laboratory, the Laboratory of the Government Chemist, the National Physical Laboratory, the Met Office and the Defence Evaluation and Research Agency. The authors acknowledge that resistance to reform proved problematic in some sectors (the old Ministry of Agriculture, Fisheries and Food "proved itself remarkably adept at avoiding the privatisation issue"), but they discover a sense of relief in some laboratories that found themselves free of the restrictive practices of the past.

The authors are much less sanguine, however, about the long-term impact of reform, in terms of the quality of science, the way that technologies are provided to government,

and in particular the transparency of reporting mechanisms and accountability to departments. Accountability to departments is now done in private, based on unpublished contracts — hardly a mechanism to instil confidence in laboratories whose functions are supposed to be transparent and carried out in the national interest.

Turning to the future, the authors express concerns resulting from the earlier policies of an "energetically reforming government" that "decided it could change things — everything in fact, including science". The dissolution of GREs that had evolved over a long period of time in response to the government's need for science and technology, the creation of organizations more by accident than design, and the cost to the Exchequer of reforming GREs that were deemed eminently non-marketable — these are topics that the jury will debate for some time.

It would have been informative to evaluate in greater depth the parallel impact of NPM reform on research-council institutions, and the spillover effect in universities, too often portrayed here as remaining, in the words of economist Robert Merton, "a public-good activity compared to a commercial one". However, this is an important and well researched book, and we should be grateful that it has been written, not least because any government bent on further reform, if it is to bring about beneficial change, would benefit from reading this recent history of the governance of science. ■

Brian Heap is an honorary fellow, St Edmund's College, University of Cambridge, Cambridge CB3 0BN, UK.

Erratum

The Surinam toad that appeared alongside the recent review by David Lodge of George W. Cox's book *Alien Species and Evolution* (*Nature* **432**, 276–277; 2004) was of the wrong species. The pictured toad (*Pipa pipa*) resides happily in South America and is not the pest (*Bufo marinus*) that has been introduced into Australia. *Nature* apologises for this mistake.

Divide, conquer and unify

Fluid dynamics: Ludwig Prandtl's ideas brought hydraulics and hydrodynamics together to found a new field.

Roddam Narasimha

In August 1904 Ludwig Prandtl, a 29-year-old professor of mechanics at the Technical University of Hanover, presented a remarkable paper at the Third International Mathematical Congress in Heidelberg. The paper was a scientific time bomb — it made no great impact at the congress, and was not translated into English until 1928. But by the 1920s and 1930s, the powerful idea in that paper and the reputation of its author had spread across the world, helping to create modern fluid dynamics out of ancient hydraulics and nineteenth-century hydrodynamics.

In retrospect, the paper, with twelve photographs of water flow past bodies, ten other diagrams and only eight equations, can be seen as full of understated daring. The heart of the paper is in one paragraph, sandwiched between some preliminary mathematics and a terse but wide-ranging exposition of the explanatory power of the idea. That exposition covered how vortices emerge in mixing layers, behind cylinders, at the edge of a plate moving normal to its plane and so on, and of when and why flows separate from the solid surface they are supposed to follow.

The tone of that key paragraph is deceptively casual. The formal problem it tackled is the flow past an aligned flat plate — a problem that is trivial in inviscid hydrodynamics, but was crucial to the new fluid dynamics that was being created. First, a small parameter, ϵ , is identified to represent viscosity, or more precisely a reciprocal of the Reynolds number. Classical perturbation methods would not work here, for the limit $\epsilon \rightarrow 0$ is singular: the limit of the full solution (for example, at the surface) is not the solution of the limiting equation. Prandtl's method was to divide the flow into different regions in which the dynamics are different — an outer inviscid flow, and a thin inner viscous ('boundary') layer next to the surface. In each of these regions only the dominant terms in the Navier–Stokes equations were collected. The equations for each region were separately solved, but their boundary conditions were cleverly selected so that the solutions blended into each other to yield a 'unified' approximation.

In that same paragraph, Prandtl also showed how the partial differential equations that governed the boundary layer could be reduced to one ordinary differential equation using a combination of the independent



Extraordinary insights: Ludwig Prandtl in 1936.

variables, and thus solved the resulting non-linear equation (which was not even explicitly written out) numerically, and provided a rough value for the drag of the plate and a sketch of the solution. And he did all this using only roughly 25 lines!

This potentially precise calculation of laminar-flow flat-plate drag opened a door between the previously sealed chambers of hydraulics and hydrodynamics. Practitioners of these disciplines had long poured scorn on each other — hydraulics was often called a science of variable constants, and hydrodynamics the mathematics of dry water. Reality, Prandtl demonstrated, was not only within the scope of Navier–Stokes equations, but even accessible to mathematics.

The idea was subsequently developed and extended, mainly through the pupils who flocked to work with Prandtl in Göttingen, where the mathematician Felix Klein helped him to establish a renowned centre for research in fluid dynamics. And the theory showed excellent agreement with experiment. Yet, astonishingly, it was regarded as no more than a clever and successful approximation or 'model' for nearly five decades. It was eventually the work of Paco Lagerstrom and his colleagues at Caltech that distilled a systematic 'method of matched asymptotic expansions' (MAX for short) as the mathematical idea inherent in Prandtl's work. In the simplest case, MAX involves studying the limiting solutions as $\epsilon \rightarrow 0$

separately in inner and outer regions, making in each of them appropriate asymptotic expansions in so-called inner and outer variables, matching the two expansions by equating (to appropriate order) the outer limit of the inner expansion with the inner limit of the outer expansion, and finally constructing a unified ('composite') solution that is everywhere asymptotic to the exact solution.

That formal mathematization of solving a whole class of singular perturbation problems had several important sequels. First it was discovered that many puzzles (from the mathematician Jean-le-Rond d'Alembert's famous paradox to others, including low Reynolds-number Stokes-type hydrodynamics) yielded to MAX. Then it turned out that the many other problems that Prandtl had brilliantly solved — such as finite wings in aerodynamics — had essentially been based on use of MAX too. When Milton van Dyke wrote the first book on the subject he crowned MAX with the adjective 'rational' — distinguishing it from the numerous non-rational, more or less ad hoc, approximations and theories that then abounded in fluid dynamics.

Werner Heisenberg said that Prandtl had "the ability to see the solution of equations without going through the calculations". Prandtl demurred, "No, I strive to form in my mind a thorough picture... the equations come only later when I believe I have understood... [and are] good means of proving my conclusions in a way that others can accept." His papers have a simplicity and directness born of supreme self-confidence. They do not trumpet their success or criticize others, but just get on with solving the central problems using all the tools available — observation (plenty of it), mathematics, calculation and modelling. Prandtl's methodological eclecticism set the style of fluid dynamics research in the twentieth century. No wonder G. I. Taylor called him 'our chief', and helped nominate Prandtl for a Nobel prize he never won. ■

Roddam Narasimha is at the Jawaharlal Nehru Centre for Advanced Scientific Research, Jakkur PO, Bangalore 560 064, India.

FURTHER READING

Birkhoff, G. *Hydrodynamics: A Study in Logic, Fact and Similitude* (Princeton Univ. Press, 1960).
van Dyke, M. *Perturbation Methods in Fluid Mechanics* (Academic Press, 1964).
Vogel-Prandtl, J. *Ludwig Prandtl* Mitt. 107, (Max-Planck-Institut für Strömungsforschung, 1993).

Electrons frozen in motion

Henrik Stapelfeldt

A single electron cloud in molecular nitrogen has been photographed. The snapshot is recorded so rapidly that it might become possible to image electron clouds as they change during fundamental molecular processes.

In medical tomography, the three-dimensional shape of a person's innards is derived from a series of two-dimensional X-ray images recorded at different angles. On page 867 of this issue, Itatani *et al.*¹ show how a new version of tomography, using laser pulses instead of X-ray beams, can determine the three-dimensional structure of a much smaller object: one of the electron clouds, or orbitals, in a nitrogen molecule (N_2).

The tomographic reconstruction is based on a series of images of the orbital shadow in different directions. It is obtained by irradiating a molecule, fixed in space, from different angles with an intense ultrashort laser pulse. A remarkable feature of the technique is that the laser pulse records the orbital image in less than a femtosecond (10^{-15} seconds). This time is so short that electrons in motion are frozen in the tomographic snapshot. In the near future, therefore, it should be possible to watch directly how electron clouds — that is, the bonds of molecules — change during chemical reactions. This would be progress indeed, and provide insight into one of the most fundamental steps in chemistry.

But how can a pulse from an optical laser image an electron cloud that is more than a thousand times smaller than the laser's wavelength? The answer lies in recent advances in understanding the interaction between molecules and intense laser pulses. The strong electric field of the laser pulse forces the most loosely bound electron away from the molecule (Fig. 1). When the field reverses its direction, half an optical period later (about 1.3 femtoseconds for the pulse used by Itatani *et al.*), the electron is forced back to collide with its parent molecule. The electron gains so much energy from the laser field that its wavelength at the moment of recollision is about 0.14 nanometres — short enough to probe the structure of the electron cloud through self-diffraction on the molecule.

It might seem most obvious to record the diffracted electrons. Instead, the authors measure the spectrum of the extreme 'overtones' of the laser light emitted in the spectral range between the ultraviolet and the X-ray regime. These overtones result from the recollision process and have been termed 'high harmonics'². The high harmonic spectrum turns out to be extremely useful because, as the authors show, the orbital structure is mapped onto the spectrum.

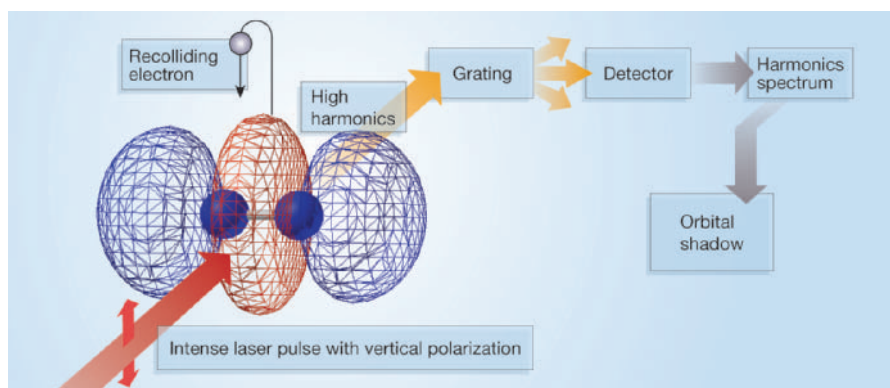


Figure 1 Laser imaging of the three-lobed electron cloud (orbital) of molecular nitrogen. Within a single optical period (2.7×10^{-15} seconds) of the laser pulse, its electric field first drives an electron away from the nitrogen molecule, along the direction of polarization (here vertical), and then forces it back to recollide with the molecule. The recollision probes the structure of the orbital and maps the orbital shadow onto the spectrum of high harmonics (overtones) of the laser frequency. The three-dimensional shape of the orbital is tomographically reconstructed from a large number of shadows measured at different angles around the molecule. The laser pulse that fixes the molecular orientation is not shown.

More precisely, for a given angle between the molecule and the direction of the recolliding electron, the shadow of the orbital can be extracted from the harmonics spectrum. The experimental strategy is to record the harmonics spectrum for a large number of collision angles. The orbital's three-dimensional shape is then reconstructed from all the corresponding shadows using a mathematical procedure that is essentially identical to that used in medical tomography.

To ensure a well-defined angle between the molecule and the recolliding electron, the spatial orientation of the molecules is fixed at the time when the intense laser pulse arrives — just as a person is fixed with respect to the direction of the X-ray beam in medical tomography. This is achieved by prealigning the molecules along the polarization of a linearly polarized alignment laser pulse, sent at a carefully chosen instant just before the intense laser pulse³. The collision angle is now controlled simply by rotating the polarization of the alignment pulse relative to the intense-pulse polarization, because the intense-pulse polarization determines the trajectory of the recolliding electron (Fig. 1).

The cloud of the valence electrons in molecular nitrogen is a fundamental quantum object found in any introductory textbook on physical chemistry. To have reconstructed it is an impressive achievement. But the real importance of Itatani and colleagues'

technique is not that it can image stationary orbitals, but its potential to provide three-dimensional snapshots of electron clouds as they undergo changes — as, for instance, in the course of chemical reactions. Experimentally, this could probably be done by preceding the intense laser pulse with a pump pulse that initiates a dynamical process in the molecular system under study. Although the principle of this approach seems similar to that of another technique, optical femtosecond laser spectroscopy, it is fundamentally different because the probe step is electron diffraction rather than absorption of an optical pulse.

In general, the work of Itatani *et al.* is part of an international effort to extend conventional femtosecond laser spectroscopy by developing femtosecond sources of short-wavelength radiation (electrons or X-rays). The object is to achieve direct, time-resolved imaging, through diffraction, of atomic positions and electron clouds in molecules and solids^{4,5}. Among these emerging techniques, Itatani and colleagues' method is unique, in that the probe beam does not originate from an external source but is generated inside each molecule. This offers unprecedented time resolution, in principle well below one femtosecond, potentially allowing studies not just of chemical reaction dynamics but also of the much faster motion of single electrons as they orbit nuclei⁶. The

next challenge is to control and use electron recollision in systems beyond those of molecules consisting of a few atoms. If that challenge is overcome, the tomography method could set new standards for information about ultrafast processes in molecules. ■

Henrik Stapelfeldt is in the Department of Chemistry, University of Aarhus,

Langelandsgade 140, DK-8000 Aarhus C, Denmark.
e-mail: henriks@chem.au.dk

1. Itatani, J. *et al. Nature* **432**, 867–871 (2004).
2. Brabec, T. & Krausz, F. *Rev. Mod. Phys.* **72**, 545–591 (2000).
3. Stapelfeldt, H. & Seideman, T. *Rev. Mod. Phys.* **75**, 543–557 (2003).
4. Stolow, A. & Jonas, D. M. *Science* **305**, 1575–1577 (2004).
5. Siwick, B. J. *et al. Science* **302**, 1382–1385 (2003).
6. Kienberger, R. *et al. Nature* **427**, 817–821 (2004).

Transmissible spongiform encephalopathies

Scrapie control under new strain

Matthew Baylis and K. Marie McIntyre

Sheep believed to be resistant to scrapie are succumbing to atypical infections and a newly identified strain of the disease. Eradication programmes based on selective breeding should be reappraised.

Scrapie is a fatal brain disease of sheep and goats and is part of the transmissible spongiform encephalopathy (TSE) family. This family also includes chronic wasting disease of deer, bovine spongiform encephalopathy (BSE) in cattle and variant Creutzfeldt–Jakob disease in humans. Unlike these relative newcomers, however, scrapie has been present for more than 250 years in European sheep flocks and has spread to many other parts of the world¹.

In the 1990s, a proportion of sheep were found to be genetically resistant to scrapie² and, consequently, many countries have established programmes to create disease-resistant national flocks by selective breeding. This new and unusual approach to disease control needs urgent reappraisal, however, as recent discoveries in Europe suggest that the selected sheep might be susceptible to strains of the disease that arise in the future. Atypical forms of scrapie have been reported in 'resistant' sheep in Germany³, Portugal⁴ and France⁵. And a recently discovered strain of scrapie, called Nor98, is now being detected in several European countries. Writing in *Journal of General Virology*, Mouton and co-workers⁶ show that this strain primarily affects sheep that rarely succumb to conventional scrapie; conversely, typically 'susceptible' sheep seem to be unaffected.

The European Union (EU) is aiming to eradicate scrapie from its member states with a combination of two approaches. The first, which has been compulsory since October 2003, involves eliminating the disease from currently infected flocks by culling entire flocks, or by genetic testing and culling of susceptible animals (Commission Regulation 1915/2003). The second approach, which will become compulsory from April 2005, has the goal of reducing the chance that uninfected flocks will become infected in the future (Commission Decision 2003/100/EC). As a minimum, rams intended for breeding in scrapie-free flocks of 'high genetic merit' must be genetically tested, and those whose

genes suggest a susceptibility to scrapie will be culled, despite being uninfected. This approach targets rams because the relatively small number used for breeding provides a fast means of spreading scrapie-resistant genes through the sheep population. Benefits are expected to filter gradually into flocks of lower genetic merit.

In some countries, broadly similar control programmes have been in operation for several years. For example, in July 2001 the UK government launched the National Scrapie Plan for Great Britain⁷, which, until the commission regulations came into force, used the same two-faceted approach but on a voluntary basis. In the United States, however, genetic testing is used to eliminate disease in scrapie-infected flocks, but breeding programmes involving scrapie-free flocks do not have to be based on scrapie genetics¹. It's important to get the strategy right: national programmes to create scrapie-resistant flocks require the continued goodwill of the

sheep industry, and are expensive. For example, more than one million UK sheep have been genetically tested, at a cost of £19 million (US\$37 million), during the first three years of Britain's National Scrapie Plan⁸, and £27.5 million has been allocated for this by the UK government for the next three years.

Whether or not a sheep can succumb to scrapie is determined by the sequence of amino acids that form its prion protein, which is encoded by its *PrP* gene. Five variants (called haplotypes) of the *PrP* gene are known to affect susceptibility. Sheep that inherit the ARR haplotype from both parents (ARR/ARR) have the greatest resistance to conventional scrapie². Conversely, sheep inheriting the VRQ haplotype from both parents (VRQ/VRQ) are extremely susceptible² to conventional scrapie and, in some infected flocks, all such animals would be expected to die from the disease. Fifteen types of sheep are definable on the basis of the pair of *PrP* haplotypes inherited from their parents; the extent to which they succumb to scrapie, as reported in the United Kingdom, is shown in Figure 1. Commission Decision 2003/100/EC aims to increase the frequency of the ARR haplotype and eliminate the VRQ haplotype.

A fundamental assumption of selective breeding programmes is that 'resistant' sheep really are, and will remain, resistant to scrapie. But this assumption is now being challenged by the new findings^{3–6} emerging from mainland Europe. In 2002, many EU member states began large-scale testing of sheep in abattoirs for scrapie, using BSE-detection methods. This surveillance has, inevitably perhaps, revealed that scrapie is more widespread than previously believed. It has also uncovered 'atypical' types⁹. These differ from conventional scrapie in several ways, such as the pattern of deposition in

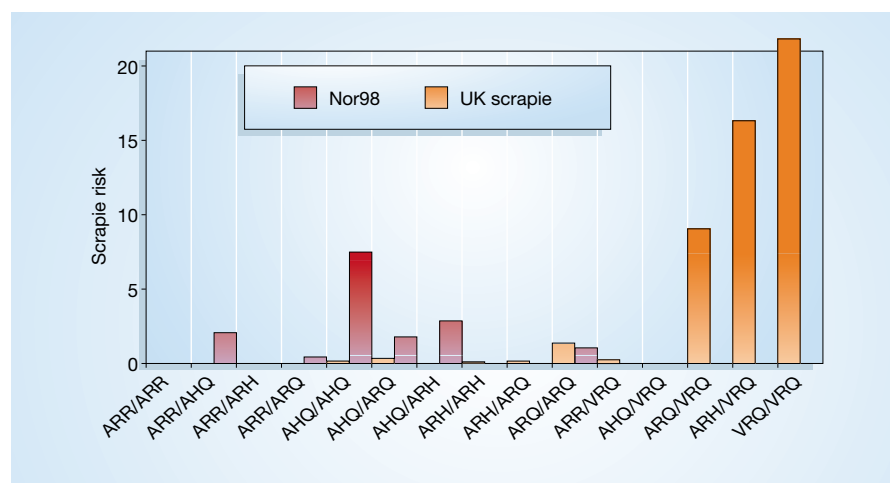


Figure 1 Scrapie risks. Variations in the *PrP* gene determine the susceptibility of sheep to scrapie. In the United Kingdom, sheep with the VRQ variant (haplotype) are most at risk of succumbing to conventional (UK) scrapie. These sheep appear, however, to be resistant to the new Nor98 strain, from which sheep with the AHQ haplotype are instead most at risk. For each genetic type, 'risk' is the proportional representation in reported scrapie cases, divided by the proportional representation in the national flock (UK) or flock mates (Nor98). Data are from refs 1, 6.

the brain of the abnormal form of the prion protein.

Importantly, a few of the atypical cases are in 'resistant' sheep. Thus, Buschmann and co-workers³ reported two atypical cases of scrapie in sheep of the ARR/ARR type in Germany, and Orge and co-workers⁴ describe a third in Portugal. Three further cases have been reported in France⁵. These discoveries confirm, under natural conditions, a finding made in sheep experimentally inoculated with BSE: ARR/ARR sheep are not, after all, totally resistant to TSEs¹⁰.

These atypical scrapie cases show some similarity to a new strain of scrapie, first detected in Norway in 1998 (ref. 11). Mowm and co-workers⁶ have now genetically tested all 38 Norwegian Nor98 cases, as well as the unaffected sheep in the flocks from which they came, allowing the relative susceptibility or resistance of each genetic type to be estimated. The results transform our understanding of the relationship between PrP haplotype and scrapie. Sheep that are the most susceptible to conventional scrapie appear to be unaffected by Nor98 (Fig. 1). Conversely, susceptibility to Nor98 is linked to the AHQ haplotype, which is generally associated with resistance (or, at most, low susceptibility) to conventional scrapie. Nor98 is also being identified over a wider area in Europe. It has been reported in Norway, Sweden, Finland and, in November 2004, in Ireland¹² and Belgium¹³.

Encouragingly, there have been no cases among ARR/ARR sheep, and so it is possible that national flocks of this genetic type may be resistant to both conventional and Nor98 scrapie. A warning shot has, however, been fired. The existence of Nor98 shows that sheep largely resistant to 'known' strains of scrapie might be highly susceptible to a novel strain identified in the future; and the discovery of atypical infections in ARR/ARR sheep presses the point that we cannot exclude these sheep from this risk. Larger epidemics of scrapie tend to be associated with flocks that have high proportions of sheep susceptible to the local scrapie strain. Accordingly, as the ARR haplotype increasingly dominates European sheep populations, there is a danger that epidemics of a strain able to attack this haplotype could be larger than those the region has experienced in the past.

One final point: the existence of the VRQ haplotype has been enigmatic, as it is not known to confer any benefits, but is often fatal for sheep exposed to conventional scrapie. Why has it not been eliminated from sheep populations by natural selection? The findings of Mowm *et al.*⁶ raise the possibility that this haplotype exists today because it conferred resistance to past strains. If so, this is a strong argument for the preservation of haplotype diversity in sheep populations, and the long-term control of scrapie

in Europe may be better achieved by the US approach: eliminating infection within infected flocks and their contacts, while preserving haplotype diversity at a national scale.

Matthew Baylis and K. Marie McIntyre are at the Institute for Animal Health, Compton, Newbury, Berkshire RG20 7NN, UK.

e-mail: matthew.baylis@bbsrc.ac.uk

1. Detwiler, L. A. & Baylis, M. *Rev. Sci. Tech. OIE* **22**, 121–143 (2003).
2. Hunter, N. *Trends Microbiol.* **5**, 331–334 (1997).
3. Buschmann, A. *et al. J. Gen. Virol.* **85**, 2727–2733 (2004).
4. Orge, L. *et al. J. Gen. Virol.* **85**, 3487–3491 (2004).
5. *Report on the Monitoring and Testing of Ruminants for the*

Presence of Transmissible Spongiform Encephalopathy (TSE) in the EU in 2003, Including the Result of the Survey of Prion Protein Genotypes in Sheep Breeds. Rep. 04-D-420525 (European Commission Health & Consumer Protection Directorate-General, Brussels, 2004).

6. Mowm, T. *et al. J. Gen. Virol.* doi:10.1099/vir.0.80437-0 (2004).
7. *National Scrapie Plan for Great Britain. Schemes Brochure* (Department for Environment, Food and Rural Affairs, London, 2001).
8. *National Scrapie Plan. Consultation on the Strategic Review* (Department for Environment, Food and Rural Affairs, London, 2004).
9. Buschmann, A. *et al. J. Virol. Methods* **117**, 27–36 (2004).
10. Houston, F. *et al. Nature* **423**, 498 (2003).
11. Benestad, S. L. *et al. Vet. Record* **153**, 202–208 (2003).
12. Onnasch, H. *et al. Vet. Record* **155**, 636–637 (2004).
13. De Bosschere, H. *et al. Vet. Record* **155**, 707–708 (2004).

Cell biology

Iron thievery

Jonathan Barasch and Kiyoshi Mori

Bacteria have many ways of stealing iron from the organisms they infect. But this thievery is not one-sided, and a newly discovered device in the mammalian tool kit does a good job of keeping bacteria in check.

A fierce battle rages between man and microbe over iron. The breakthrough reported by Flo and colleagues¹ on page 917 of this issue exemplifies the lengths to which microorganisms must go to obtain iron — and the remarkable ingenuity of our own natural countermeasures for denying them this essential metal. So far, humans are winning, and we're doing it with a member of an obscure family of proteins called the lipocalins.

Iron is a problem for all that live by it. This is because the ferrous ion, although soluble in biological fluids, generates toxic hydroxyl radicals in the presence of oxygen. The oxidized (ferric) form is better behaved, but it is insoluble in water at neutral pH (ref. 2), producing ferrihydroxide precipitates; ferriphosphate precipitates also form with the biologically abundant phosphate ion. It is somewhat paradoxical that these iron ions are so inaccessible, given that iron is the key catalytic site of many of the enzymes and gas-transporting proteins in cells.

To solve the problem of taking up iron into cells, organisms have developed numerous iron-solubilization technologies. In mammals, birds, fish and amphibians, a highly conserved protein called transferrin is needed to move the major load of extracellular iron into cells. Transferrin binds extracellular iron with high affinity (10^{-20} M), docks at transferrin receptors on the cell membrane, and is taken up into cells by means of the invagination of sections of this membrane. The iron is then unloaded from transferrin into specialized intracellular compartments, from where it can be transferred to the cytoplasm.

In bacteria and fungi, meanwhile, other iron-capturing tools are brought into play.

For instance, these organisms can secrete and then reclaim a bewildering array of iron-avid peptides and ester derivatives, known collectively as siderophores³. These are classified as phenolates–catecholates or hydroxamates, or mixtures of the two forms. Enterochelin, a major catecholate siderophore, chelates iron with extraordinarily high affinity (10^{-49} M)⁴, docks at unique receptors — the Fep proteins — on the bacterial surface, and is destroyed by esterases after being internalized.

So what happens when bacteria grow in a host that also covets iron (mammalian blood serum has just 10^{-26} M free iron⁵), or when two microorganisms compete for the same source of metal? Essentially, thievery reigns. For instance, when in competition with fungi, bacteria can hijack iron that is targeted to the fungi by producing receptors for fungal siderophores that the bacteria themselves do not make. When in competition with animal hosts, bacteria can expropriate transferrin (or the related lactoferrin) wholesale, or simply remove its iron by using siderophores. Mycobacteria use the siderophore carboxymycobactin⁶ to strip iron from mammalian ferritin, another iron-binding protein. And fungi of the genus *Rhizopus*⁷ have been reported to capture an iron-chelating medicine, deferoxamine, allowing these organisms to multiply to levels that are disastrous for the host. Perhaps not surprisingly, giving iron to patients⁸ (and, as Flo *et al.* show¹, to mice) worsens the outcome of bacterial and mycobacterial infections.

But the hosts have a few tricks of their own. If they hoard iron, they can control the microbial population. One way of doing this is to reduce the amounts of transferrin. Iron recycling from macrophage cells (a type of immune cell) can also be shut down. And



100 YEARS AGO

Earthquakes. By Clarence Edward Dutton. Major Dutton's work belongs to another category, and rather than telling us what earthquakes do, his main object has been to tell us what they are... Everything is discussed with a minimum of mathematics from a strictly scientific standpoint, whilst that which is sensational has properly been most carefully put under taboo. A justification for the exclusion of what is of practical importance, which gives not only to the man in the street but to Governments some inkling as to the use of earthquakes, is not so apparent. It is extremely likely that a Prime Minister may not care a twopenny-bit whether the inside of the world on which he lives is red hot stone or cold, while he might be extremely interested to know that seismograms may afford a satisfactory explanation for the interruption of his cablegrams. The importance of earthquake writings to communities who have been alarmed by accounts of disasters in foreign countries is self-evident, while it would at least be consoling to those who were suddenly cut off from the outer world by the failure of their cables to learn whether such failures were the result of an operation of war or of nature.

From *Nature* 15 December 1904.

50 YEARS AGO

The structure of fibrous proteins has long been a subject of controversy. X-ray and electron microscope evidence has accumulated which suggests that single chains may not run the whole length of the fibril, but that the latter is made up of an aggregation of smaller parts of quite definite size. In collagen the sub-unit has been considered to be a protofibril of size about 640×12 A., although recently Schmitt has proposed a unit of about 2000×50 A., which he has named 'tropocollagen'. Striations of axial lengths about 210 A. (particularly in developing material) and 70 A. are also observed in electron micrographs of collagen. It is of interest to note that evidence for structure of size approximately 200 A. is found in α -keratin and 230 A. in fibrin, although the recurrence of this figure may be no more than coincidental. We have obtained X-ray diffraction evidence from dry collagen fibres which also suggests that the predominant 640 A. period is divided into sub-units of length about 210 A. A. C. T. North, P. M. Cowan, J. T. Randall

From *Nature* 18 December 1954.

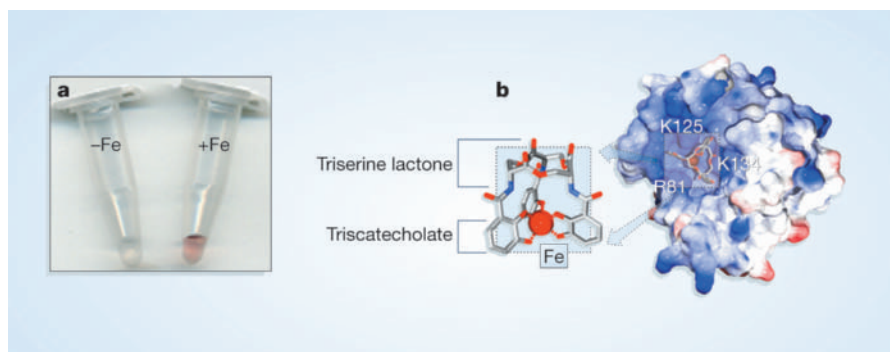


Figure 1 Snatching bacterial iron. **a**, In the presence of iron, solutions of the lipocalin 2 protein, synthesized in the laboratory, are distinctively *rosé*; without iron, they are colourless¹⁰. This demonstration led to the discovery that a ligand for lipocalin 2 is bacterial enterochelin — a secreted iron-chelating molecule. **b**, Bacterial enterochelin is composed of a triseric lactone molecule, which has three catecholate groups. These groups, between them, chelate an iron atom or ion. The iron-enterochelin complex can in turn be buried in the calyx of lipocalin, between three positively charged amino acids (R81, K125 and K134, where R is arginine and K is lysine). Flo *et al.*¹ have now found that lipocalin 2 is necessary for mice to keep bacterial infections in check; it works by stealing bacterial iron. (Molecular models courtesy of R. K. Strong, Fred Hutchinson Cancer Research Center, Seattle.)

Flo *et al.* reveal yet another mechanism: mammals can steal siderophores.

The authors first discovered that, during a bacterial infection, the mammalian liver, spleen and macrophages synthesize the lipocalin 2 protein, raising its concentration in the serum by log orders of magnitude. The lipocalins are a large family⁹ that have attracted structural biologists' attention because, although they vary in amino-acid sequence, their three-dimensional structures are remarkably conserved. They are made up of strands of so-called β -sheets, which form a barrel or a cup-like structure, or calyx, that carries small chemicals. A well-known example is the retinol-binding protein, with its retinoic acid ligand. So what is the chemical carried by lipocalin 2? Two years ago, Goetz *et al.*¹⁰ showed that iron-containing solutions of cloned lipocalin 2 are bright red (Fig. 1). Using several structural-analysis techniques, they tracked down the source of the red colour to the presence of iron-bound enterochelin.

The affinity of lipocalin 2 for enterochelin is very high (10^{-10} M), suggesting that this bacterial siderophore might be the authentic ligand (indeed, lipocalin 2 has been tentatively renamed siderocalin¹⁰). The idea is not so far-fetched — nitrophorin lipocalins (from the salivary gland of *Rhodnius prolixus*, the insect that spreads the parasite that causes Chagas' disease) were already known to carry iron-loaded haem groups¹¹. And lipocalin from human tears binds to bacterial and fungal siderophores¹².

Nonetheless, all of this might have been dismissed as an artefact of the cloning of lipocalin 2. Instead, it seems that it actually revealed an *in vivo* function. Using mice in which lipocalin 2 had been knocked out, Flo *et al.* unequivocally show that the protein is essential for limiting the growth of common bacteria that produce enterochelin, but does

not affect the growth of bacteria that rely on non-catecholate siderophores or other methods of iron acquisition. Eliminating lipocalin 2 increased bacterial colony counts by log orders of magnitude, and caused the rapid demise of the knockout mouse. Conversely, lipocalin 2 limited bacterial growth *in vitro* and *in vivo*, until reversed by excess siderophore. So lipocalin 2 directly regulates iron-dependent bacterial proliferation in mice — and probably, therefore, in humans.

These findings raise many issues for us to ponder. For instance, what happens to lipocalin 2 once it has chelated the siderophore and its iron? Recent work suggests¹³ that the kidneys filter and reabsorb the complex, recycling its iron. This pathway protects the kidney tubules from stress, hinting that the function of lipocalin 2 does not stop with stealing iron from microbes — the iron-loaded protein may also have other activities in mammalian cells. Indeed, lipocalin 2 is produced in massive amounts by cells that are damaged by chemicals or oxygen depletion in the absence of bacterial infection. This suggests that the production of lipocalin 2 is either a stress response that was originally intended as a defence against bacteria, or perhaps a mechanism for shuttling an unidentified endogenous ligand into cells. That endogenous ligand might even be a mammalian version of the bacterial siderophore, although the existence of such molecules has not been confirmed¹⁴.

Another possibility raised by the new findings is that many other lipocalins, whose functions are still unknown, are also iron carriers and provide an innate defence against microorganisms. And perhaps our most dangerous bacterial enemies, such as *Pseudomonas*, escape our control by making siderophores that are beyond the reach of our lipocalin-based surveillance mechanisms. Flo and colleagues' results¹ are an

open invitation to researchers to custom-design not new artificial iron chelators, but new siderophore chelators.

Great resourcefulness is required to capture insoluble iron. In the fight between mammals and microbes for this metal, it seems that, thanks to lipocalin 2, mammals are the winners — at least for now.

Jonathan Barasch is in the Departments of Medicine and of Cell Biology, and Kiyoshi Mori is in the Department of Medicine, Columbia University, New York City, New York 10032, USA.

e-mails: jmb4@columbia.edu

km2112@columbia.edu

1. Flo, T. H. *et al. Nature* **432**, 917–921 (2004).

2. Klausner, R. D. & Rouault, T. *Harvey Lect.* **92**, 99–112 (1996–97).
3. Neilands, J. B. *J. Biol. Chem.* **270**, 26723–26726 (1995).
4. Loomis, L. D. & Raymond, K. N. *Inorg. Chem.* **30**, 906–911 (1991).
5. Otto, B. R. *et al. Infect. Immun.* **59**, 2999–3003 (1991).
6. Gobin, J. & Horwitz, M. A. *J. Exp. Med.* **183**, 1527–1532 (1996).
7. Boelaert, J. R. *et al. J. Clin. Invest.* **91**, 1979–1986 (1993).
8. Ratledge, C. *Tuberculosis* **84**, 110–130 (2004).
9. Flower, D. R., North, A. C. & Sansom, C. E. *Biochim. Biophys. Acta* **1482**, 9–24 (2000).
10. Goetz, D. H. *et al. Mol. Cell* **10**, 1033–1043 (2002).
11. Montfort, W. R., Weichsel, A. & Andersen, J. F. *Biochim. Biophys. Acta* **1482**, 110–118 (2000).
12. Flückinger, M., Haas, H., Merschak, P., Glasgow, B. J. & Redl, B. *Antimicrob. Agents Chemother.* **48**, 3367–3372 (2004).
13. Mishra, J. *et al. J. Am. Soc. Nephrol.* **15**, 3073–3082 (2004).
14. Yang, J. *et al. Mol. Cell* **10**, 1045–1056 (2002).

Archaeology

Much rowing for fish

Daniel Pauly

A thousand years ago, there was a shift in the fish diet in England from freshwater to marine species. The relevant case history, derived from picking through leftovers, has a contemporary resonance.

Aelfric (AD 955–c. 1020), the Abbot of Eynsham, near Oxford, has in his *Colloquy* a master asking a fisherman why he does not fish in the sea. The fisherman answers, “Sometimes I do, but rarely, because it is a lot of rowing for me to the sea”¹. Yet, in spite of all the rowing this transition required, Aelfric lived during the period when the fish that the English ate changed from mainly freshwater species to mainly marine ones — principally herring (*Clupea harengus*) and representatives of the cod family, the ‘gadids’, with cod (*Gadus morhua*) dominating. We know this from a study by Barrett and colleagues, just published in *Proceedings of the Royal Society*², which documents the contents of 127 ‘assemblages’ (of old fish bones) from settlements around England, covering a period from the seventh to the sixteenth century.

There is a large community of archaeologists and historians working up the material that was left over (in middens, old latrines and so on) from the meals taken by our ancestors. But this literature seems rarely to be used to draw inferences about the historical exploitation of natural resources, at least in the fisheries literature, which is a pity. Using such remains, Barrett *et al.* have identified the origin of intensive marine fishing, and hence a period that should serve as a baseline when evaluating the present state of marine ecosystems — at least in the waters around England.

The transition towards marine species, they suggest, was due to a decline in freshwater species. This decline was perhaps caused by pollution from mills and from agricultural run-off (farming was expanding during that period), and by overfishing of what are, after all, limited resources, relative to the demand

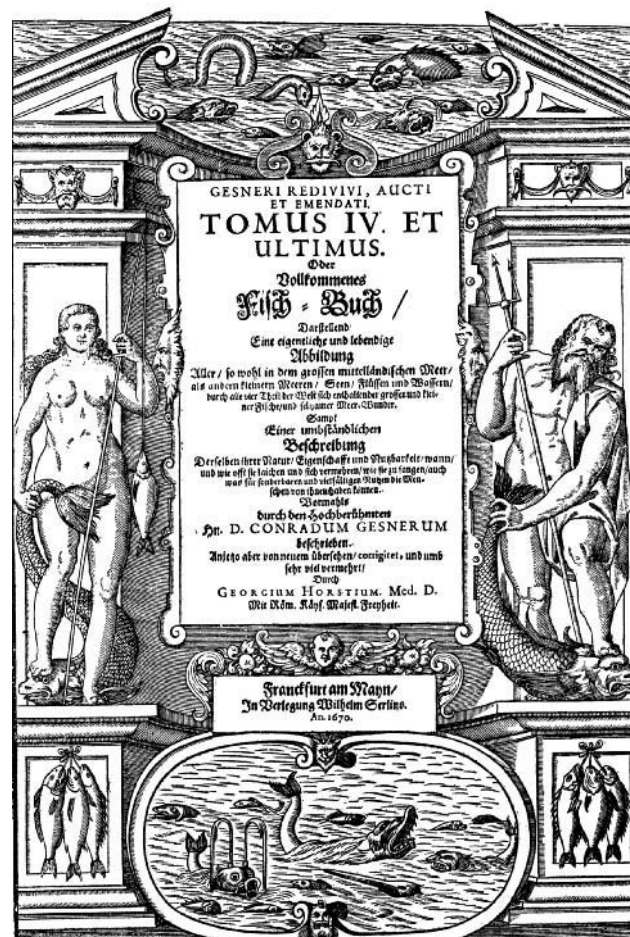
from a then-growing urban population.

Barrett *et al.* also point out that the period between 950 and 1050, which brackets the transition from freshwater to marine fishes in the English diet, was relatively warm. Such

conditions were not favourable to herring or cod in the seas around England, where both species are near the southern end of their ranges. Hence, only a massive (and historically undocumented) increase of trade with Norway, where the abundance of cod and herring does tend to increase with temperature, could link the observed dietary transition to shifting abundances resulting from environmental change.

In addition, Barrett *et al.* identify a group of ‘intermediates’ between strictly freshwater species and marine species. These are the flatfishes, which include species that migrate between marine and freshwater environments. And indeed, two flatfish species are mentioned by Aelfric’s fisherman, when he lists, upon being asked what he catches in the sea: “Herring and salmon, porpoise and sturgeon, oysters and crabs, mussels, winkles, cockles, plaice and flounders and lobsters and many similar things.”

There is another aspect of this study³ that the authors themselves don’t raise, but that I can’t resist mentioning here. This is that the analyses not only confirm and narrow down a freshwater-to-seawater transition previously known, though with less precise dates, from other parts of northwestern Europe (see ref. 1 for references), but they deepen our understanding of the last global transition towards the consumption of marine fish. This



Fish for all. Several centuries after the time of Aelfric’s fisherman, Conrad Gessner vividly described and depicted various species of European fishes in his *Fisch-Buch*, published in Frankfurt. It was part of the encyclopaedia that encompassed the zoological knowledge of his time.

transition, in a sense, sent us back to where we were tens of thousand of years ago, before the invention of agriculture, when many (most?) humans were living and migrating along coastlines, relying on shellfish and, starting from Africa, gradually reaching islands as far flung as Tasmania and Tierra del Fuego³.

These migrations, which brought our African ancestors to Europe and Asia, and thence to Australia and Oceania, and to the Americas, were in most cases followed by migrations inland, driven by hunting of larger animals, and later by agriculture. And interestingly, there is documentation, also from Britain, of a change in diet some 5,200 years ago, at the start of the Neolithic and of British agriculture, from marine animals to inland animals and plants⁴.

The transition from eating freshwater fish to eating marine fish, documented by Barrett *et al.*, is thus part of a secondary transition. Marine fisheries catches are declining worldwide⁵, and the global demand for fish is increasing. So will another transition back to freshwater fish and inland species occur? Or will we be able to 'row' even more, and maintain our present level of marine fish consumption? ■

Daniel Pauly is at the Fisheries Centre, Lower Mall Research Station, 2259 Lower Mall, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada.

e-mail: d.pauly@fisheries.ubc.ca

1. Lacey, R. & Danziger, D. *The Year 1000: What Life Was Like at the Turn of the First Millennium* (Little, Brown, New York, 1999).
2. Barrett, J. H., Locker, A. M. & Roberts, C. M. *Proc. R. Soc. Lond. B* **271**, 2417–2421 (2004). See also Electronic Appendix A: "Summary information and reference for the fish bone assemblages surveyed."

3. Erlandson, J. M. in *The First Americans: The Pleistocene Colonization of the New World* (ed. Jablonski, N. G.) 59–92, Mem. Calif. Acad. Sci. No. 27 (Calif. Acad. Sci., San Francisco, 2002).
4. http://news.nationalgeographic.com/news/2003/09/0924_030924_neolithicdiet.html#main (accessed 20 September 2004).
5. Watson, R. & Pauly, D. *Nature* **414**, 534–536 (2001).

Palaeoclimate

A balmy Arctic

Christopher J. Poulsen

Analyses of sediments retrieved from a drifting ice island suggest that the Arctic Ocean may have been ice free and as warm as 15 °C about 70 million years ago. Therein is a challenge for climate models.

Various lines of evidence show that Earth's climate was much warmer during the Cretaceous period than it is today. Yet that evidence — fossil plants and animals, sedimentary features and geochemical indicators — is sparse, spotty and often inexact, making the magnitude and distribution of Cretaceous temperatures highly uncertain. In a testament to quality over quantity, Jenkyns *et al.*¹ (page 888 of this issue) have produced a single new datum from one of the coldest spots on Earth's surface: their work clarifies the nature of extreme warmth in the Cretaceous, and

reveals that the past was radically different from the present.

Ironically, Jenkyns and colleagues' evidence for a warm Arctic climate was retrieved by drilling through an ice island drifting over the Alpha Ridge, one of three submarine ridges that divide the Arctic sea floor (see the map on page 889). Sediments deposited on the modern Arctic sea floor are oxidized, rich in silt and sand, and have only sparse evidence of planktonic life: with nearly 3 metres of sea ice capping its surface, the Arctic Ocean is not a hospitable environment for most organisms. In sharp contrast,

Planetary science

An ill solar wind

On 1 November last year, my father phoned me from Tenerife to ask why he wasn't able to receive the BBC World Service. His concern was not at missing the news bulletins, but rather at what might have happened to his beloved transmitters at Rampisham in the south of England or on Ascension Island in the mid-Atlantic (my father is a chartered electrical engineer). In fact, it was unusually high sunspot activity that was upsetting his reception. Elsewhere in this issue, Daniel Baker and colleagues report exactly how dramatic were the effects on Earth's atmosphere of the solar winds from those sunspots (*Nature* **432**, 878–881; 2004).

Short-wave radio signals are transmitted around the world by refraction and reflection from a layer of ions in the atmosphere; this Appleton layer (named after Edward Appleton, who discovered it in 1926) lies between 150 and 1,000 km above Earth's surface. Higher

still, at 3,000–6,000 km and 20,000–25,000 km above the Equator, are the Van Allen belts — two toroidal regions of high-energy ions, mainly electrons and protons, trapped by Earth's magnetic field. The relatively calm region between the belts is a future orbit of choice for artificial satellites. The inner edge of the outer belt also corresponds to the limit of Earth's plasmasphere, which is dense with ions.

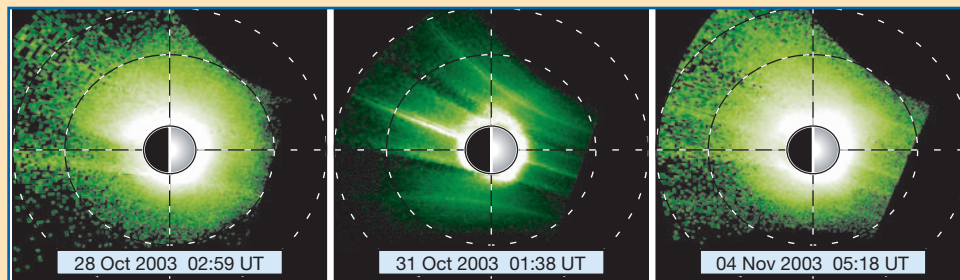
Around Halloween in 2003, two spacecraft — the Solar Anomalous and Magnetospheric Particle Explorer (SAMPEX) and the Imager for Magnetopause-to-Aurora Global

Exploration (IMAGE) — detected massive distortions in the Van Allen belts and the plasmasphere as the solar winds hit. The gap between the belts was obliterated throughout November, as electrons were 'blown' through into the inner belt, increasing its electron density 50-fold — a situation that has persisted ever since. The effect on the plasmasphere (pictured) was even more dramatic, if shorter-lived. On 28 October, it extended as usual to about 19,000 km above Earth's surface; by 31 October it had contracted to 6,000 km, and in places to as low as 3,000 km,

returning to close to normal within a few days.

These changes greatly reduced Earth's natural shields against solar and cosmic radiation. Contact was lost with satellites; astronauts on the International Space Station took cover in a heavily shielded service module; the US Federal Aviation Administration issued its first-ever radiation alert to passengers flying above 7,500 m; and the power system in Malmö, Sweden, failed.

Small wonder, then, that the Appleton layer could not relay the World Service to my father in the Canaries. **Christopher Surridge**



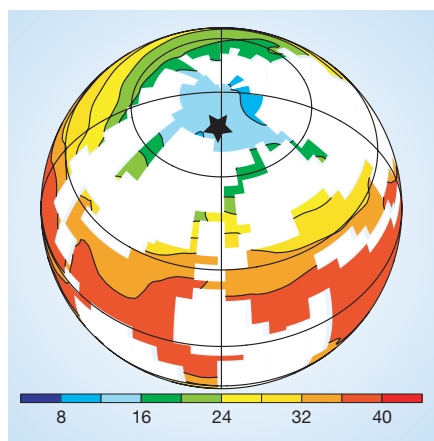


Figure 1 Average annual sea surface temperature for the Cretaceous predicted by an ocean-atmosphere climate model. Units are °C, and the white areas indicate land masses. Using data from the Alpha Ridge (star), Jenkyns *et al.*¹ conclude that during the Late and Middle Cretaceous, seawater temperatures in the Arctic were 15 °C or more. Climate models cannot simulate such warm conditions without incorporating very high CO₂ levels that exceed the estimates of CO₂ at that time⁹. In this example, Arctic seawater temperature is nearly 15 °C. But the model concentration of atmospheric CO₂ required is 7,100 p.p.m.v., 20 times present levels, and 3–6 times larger than the estimated levels for the Cretaceous.

the Late Cretaceous sediment from the Alpha Ridge consists of an ooze that is rich in diatoms (unicellular algae) and black mud that contains woody fragments, leaf cuticles, spores and pollen. The sediments are also notable for what they do not contain — dropstones, the non-marine debris that fall from the base of icebergs. All in all, these sediments were evidently the product of fertile, ice-free marine waters.

Although the sedimentary evidence from the Arctic marine cores is suggestive of warm, ice-free conditions, it is a very imprecise thermometer. For this reason, palaeo-oceanographers often draw on geochemical tools, such as the ratio of oxygen isotopes in carbonate shells from marine microorganisms, that can provide precise estimates of seawater temperature. Because of the absence of carbonate in Arctic Cretaceous sediments, traditional methods could not be used. So Jenkyns *et al.*¹ turned to a novel method (called TEX₈₆) for quantifying sea surface temperature. They examined the composition of lipids in the membranes of Crenarchaeota, floating marine microorganisms that range in size from 0.2 to 2.0 micrometres, and which have been found in sediments up to 112 million years old.

Earlier work showed that the composition of organic compounds in the membranes of modern Crenarchaeota is highly correlated with seawater temperature and may well be a biological adaptation to the

thermal state of the marine environment². The TEX₈₆ method has previously been tested on Cretaceous sediments from low latitudes, and the results are in good agreement with estimates of seawater temperature using oxygen isotopes³.

Applying this new geochemical tool to 70-million-year-old organic material from the Alpha Ridge, Jenkyns *et al.*¹ calculate that the average sea surface temperature was 15 °C. Sea water of similar temperature is presently found off (for example) Maryland and France, at latitudes between 35° N and 45° N. For a region blanketed in darkness for half of the year, the Arctic Ocean was astoundingly warm. It may have been even warmer earlier in the Cretaceous: proxy evidence indicates that the climate had been slowly cooling for nearly 20 million years. Although astounding, the new estimate of Arctic seawater temperature is not without precedent. Fossil evidence of the tropical breadfruit tree *Artocarpus dicksoni* and of champsosaurs, extinct crocodile-like reptiles, has been found in sediments from the high Canadian Arctic dating to the middle Cretaceous (90–100 million years ago)^{4,5}.

Why was the Cretaceous so warm? The most likely explanation is high concentrations of atmospheric carbon dioxide that resulted from high rates of volcanic outgassing. Atmospheric CO₂ is the primary greenhouse gas that is believed to have contributed to global warming since the beginning of the industrial revolution. In the Cretaceous, CO₂ levels were probably three to six times those of today, giving rise to a super-greenhouse climate. Yet model simulations^{6,7} of climate during the Cretaceous, using these estimates of past CO₂ levels, produce Arctic Ocean temperatures of 2–6 °C, much lower

than the estimates of Jenkyns *et al.*¹.

Although it is possible that these estimates of past CO₂ are too low, problems still remain for the climate models. Specifying higher CO₂ levels warms the polar regions (Fig. 1). But it also leads to very high temperatures at low latitude — temperatures that exceed the estimated values using proxy methods and approach the tolerance level of organisms⁸.

Why do simulations of the Cretaceous climate predict polar temperatures that are too cold and Equator-to-pole temperature gradients that are too large? The solution to the problem may lurk in the climate models themselves. Attempts have been made to solve it by incorporating the effects of ocean heat transport, stratospheric clouds, ocean passageways, and vegetation. The result, however, has been only incremental improvements. Climate models still do an inadequate job of simulating the extreme warmth of a past greenhouse world — a troubling proposition for predictions of a future greenhouse world.

Christopher J. Poulsen is in the Department of Geological Sciences, University of Michigan, Ann Arbor, Michigan 48109, USA.

e-mail: poulsen@umich.edu

- Jenkyns, H. C., Forster, A., Schouten, S. & Sinninghe Damsté, J. S. *Nature* **432**, 888–892 (2004).
- Schouten, S., Hopmans, E., Schefuß, E. & Sinninghe Damsté, J. S. *Earth Planet. Sci. Lett.* **204**, 265–274 (2002).
- Schouten, S. *et al. Geology* **31**, 1069–1072 (2003).
- Tarduno, J. *et al. Science* **282**, 2241–2244 (1998).
- Morton, J. *Fruits of Warm Climates* (Creative Resources Systems, Miami, 1987).
- Poulsen, C., Barron, E., Peterson, W. & Wilson, P. *Paleoceanography* **14**, 679–697 (1999).
- Otto-Bliesner, B., Brady, E. & Shields, C. *J. Geophys. Res.* **107**, 10.1029/2001JD000821 (2002).
- Barron, E. *Earth Sci. Rev.* **19**, 305–338 (1983).
- Bice, K. & Norris, R. *Paleoceanography* **17**, 10.1029/2002PA000778 (2002).

Cell biology

Popping out of the nucleus

André Hoelz and Günter Blobel

Previous structural snapshots have provided insight into how proteins are imported into the cell nucleus. The structure of an export complex now completes the molecular picture of the nuclear transport cycle.

A notable stage in evolution was the compartmentalization of cells, producing the eukaryotic cell — the cell type found in organisms from fungi to humans. During this process, the genetic information became enclosed inside a nucleus, spatially separated from the rest of the cell. The only way that macromolecules can be transported into and out of the nucleus is through large protein assemblies termed nuclear pore complexes, which are located in circular openings in the nuclear membrane, or ‘envelope’. Mobile proteins that ferry cargo are also required for import

and export, as is another protein, Ran, whose conformational state alters the affinity of the carriers for their cargo. On page 872 of this issue, Matsuura and Stewart¹ reveal the structure of one nuclear export protein complex. Comparing this snapshot with structures of nuclear import complexes sheds light on why Ran has different roles depending on whether cargo is being directed into or out of the nucleus.

Nuclear transport is mediated by short sequence elements in cargo molecules: cargo carrying a nuclear localization sequence (NLS) is imported, whereas a nuclear export

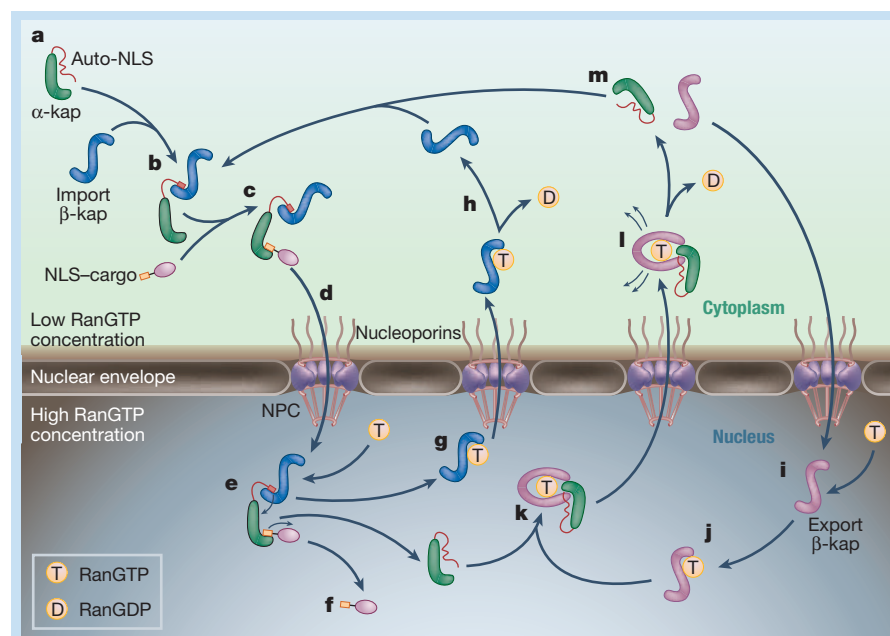


Figure 1 The nuclear transport cycle. Cargo to be imported contains a nuclear localization sequence (NLS). The α -kap protein is needed for cargo import, and is itself a cargo for export. **a**, In the cytoplasm, α -kap's NLS-binding site is blocked by its auto-NLS. This prevents a cargo's NLS from binding to α -kap. **b**, In the presence of an import β -kap, the auto-NLS of α -kap binds to β -kap, allowing α -kap to bind NLS-cargo (**c**). **d**, The ternary complex docks at nucleoporins (proteins of the nuclear pore complex, NPC) and enters the nucleus. **e**, The complex dissociates when RanGTP interacts with the amino-terminal arch of β -kap, causing the release of α -kap's auto-NLS, which in turn displaces the cargo's NLS. **f**, The cargo is thus released. **g**, The import- β -kap-RanGTP complex is exported. **h**, GTP is hydrolysed, producing RanGDP, which is released. Export of the α -kap, meanwhile, requires an export β -kap. **i**, **j**, The export β -kap forms a complex with RanGTP. **k**, This complex binds to α -kap and is exported. Matsuura and Stewart¹ find that assembly of this ternary complex stores energy by twisting the export β -kap from an S-shape into a horseshoe-like conformation, in which both arches embrace RanGTP. **l**, Arrival of the complex on the cytoplasmic face of the NPC triggers GTP hydrolysis. The complex releases the stored energy and pops open. The conformational change in β -kap liberates α -kap and RanGDP. **m**, The α -kap enters a new round of nuclear import, whereas the export β -kap may relocate to the nucleus for another export cycle.

sequence (NES) is used for export. These sequence elements are recognized by transport factors, collectively termed karyopherins (also referred to as importins, exportins or transportins), which ferry the cargo.

The karyopherins ('kaps' for short) come in two flavours: α -karyopherins (α -kaps) and β -karyopherins (β -kaps). The α -kaps bind to the NLS of a cargo protein, and function only in nuclear import. By themselves they are unable to bind to the proteins — collectively termed nucleoporins — that make up the nuclear pore complex. Therefore, they team up with β -kaps, which by binding to nucleoporins provide access to the nuclear pore complex. In general, β -kaps can bind to either an NLS or an NES, and also to nucleoporins. They can function in nuclear import (import kaps), export (export kaps), or both.

The assembly and disassembly of cargo-kap complexes is governed by Ran, a protein that comes in two conformationally distinct states depending on whether it is bound to the nucleotide guanine triphosphate (GTP) or to guanine diphosphate

(GDP)^{2,3}. Intriguingly, RanGTP has different functions in import and export: it causes the disassembly of complexes consisting of cargo plus import β -kap, but is required for the assembly of a cargo-export-kap complex. Thus, because RanGTP is primarily located in the nucleus, it favours the assembly of nuclear export complexes, but causes the disassembly of import complexes that have entered the nucleus.

Kaps are composed of helical molecular motifs called HEAT or Armadillo repeats, which are stacked on top of each other to form a highly flexible, superhelical structure. The α -kaps contain a self-inhibitory segment at one end (the amino terminus) — a segment that functions as an 'auto-NLS'. This sequence can fold back in an intramolecular reaction to occupy α -kap's NLS-binding site, which lies in a groove of the superhelix⁴.

This constellation prevents the binding of NLS-containing cargo (Fig. 1). Alternatively, the auto-NLS of the α -kap can bind to the NLS-binding site of an import β -kap (specifically, kap- β 1), allowing NLS-containing cargo to bind to the α -kap. After being transported into the nucleus, the import complex

is dissociated by the binding of RanGTP to the β -kap. This triggers the dissociation of α -kap's auto-NLS from the β -kap, and an auto-NLS-induced displacement of the cargo's NLS from the α -kap. Structural studies of the import β -kaps revealed that they can form two arches in an S-shaped configuration⁴, with the binding partners for each arch as indicated in Figure 1.

Matsuura and Stewart¹ now describe the first crystal structure of an export complex, consisting of an export β -kap (Cse1p from yeast), its export substrate (the yeast protein Kap60p) and RanGTP. Kap60p is the only α -kap of yeast, and must be exported from the nucleus in order to serve in another round of nuclear import⁴.

The structures of Kap60p and RanGTP in the context of the ternary export complex¹ are not significantly different in conformation from the structures of the two individual proteins. Strikingly, however, the two arches of the export β -kap (Cse1p) embrace both RanGTP and α -kap¹. Unlike in the import β -kaps, where only one arch of the S-shaped structure is engaged at a time⁴, the cooperative binding of both RanGTP and the export substrate forces both arches of the export β -kap from an S-shaped configuration into a more compact, horseshoe-like configuration (Fig. 1). The ternary complex is formed only when the auto-NLS of the α -kap is in place in this protein's NLS-binding site. This allows for additional contacts to be made between the auto-NLS and the export β -kap, ensuring that only cargo-free α -kap is exported. Strikingly, in this first structurally analysed case, the NES is not a contiguous sequence, but is constructed from several contact points along the export substrate.

Matsuura and Stewart's results also strongly suggest that the RanGTP-induced conformation of the export β -kap results in energy storage. When the ternary export complex arrives on the outer (cytoplasmic) face of the nuclear pore complex, it is exposed to proteins that cause the hydrolysis of the bound GTP. This releases the energy stored in the compacted export β -kap, which pops open and releases the α -kap for a new round of nuclear import. It remains to be seen whether the cooperative binding of RanGTP and other export substrates causes similar conformational changes in other export kaps.

André Hoelz and Günter Blobel are in the Laboratory of Cell Biology, Howard Hughes Medical Institute, The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA.

e-mails: hoelza@mail.rockefeller.edu

blobel@mail.rockefeller.edu

1. Matsuura, Y. & Stewart, M. *Nature* **432**, 872–877 (2004).
2. Vetter, I. R., Nowak, C., Nishimoto, T., Kuhlmann, J. & Wittinghofer, A. *Nature* **398**, 39–46 (1999).
3. Scheffzek, K., Klebe, C., Fritz-Wolf, K., Kabsch, W. & Wittinghofer, A. *Nature* **374**, 378–381 (1995).
4. Chook, Y. M. & Blobel, G. *Curr. Opin. Struct. Biol.* **11**, 703–715 (2001).

Evolutionary biology

Chromosome chain makes a link

Laura Carrel

The platypus' sex chromosomes are as peculiar as its appearance. Five X and five Y chromosomes form a remarkable chain-like configuration in male reproductive cells that ensures appropriate sperm formation.

An Australian Aboriginal legend proposes that the platypus is the offspring of a duck and a water-rat, with a duck-like bill and webbed toes, but four feet and fur like a rat. However, despite some avian and reptilian features, phylogenetic studies classify the platypus as a monotreme, an egg-laying mammal. Monotremes diverged from the rest of the mammalian lineage some 210 million years ago¹, and examining them not only gives a unique evolutionary perspective to biological processes specific to mammals, but also provides links to more distant animals, such as birds.

One example of this is provided by studies of platypus chromosomes, described in this issue (page 913)², and in *Proceedings of the National Academy of Sciences*³, by members of the same groups. The reports clarify the composition and behaviour of platypus sex chromosomes, and challenge current theories on the evolution of sex determination.

Sex chromosomes and sex-determining genes differ among organisms (Fig. 1, overleaf). In most mammals, females have two X sex chromosomes, and males have an X

and a Y. The Y chromosome harbours a sex-determining gene, *SRY*, that functions as a master switch, initiating a developmental programme that directs the embryonic gonad to form a testis rather than an ovary⁴. However, sex determination seems to have evolved independently in some organisms⁵. In birds, for example, the females are the ones with two different sex chromosomes, designated ZW, whereas males have ZZ⁶. Birds have no *SRY* gene, but a candidate testis-determining gene called *DMRT1* occurs on their Z chromosome⁶ (Fig. 1). Although monotremes might be expected to have a similar system to other mammals, they lack *SRY* and, until now^{2,3}, even the composition of sex chromosomes in the platypus was poorly understood.

Early studies of platypus male reproductive cells undergoing meiosis, the specialized cell division process that creates sperm, showed that a group of chromosomes were linked in an unusual chain-like

configuration¹. A similar meiotic structure occurs in a few disparate organisms, such as evening primrose and termite^{7,8}, but it has not been seen before in vertebrates. Grützner and colleagues^{2,3} have created fluorescent 'paints' for individual platypus chromosomes. By tagging the chromosomes in this manner, they could determine the number of copies of each chromosome in males and females, and their order in the male meiotic chain (see Fig. 2,



D. WATTS/NATUREPL.COM

Infectious disease

Past and future of an old foe

Epidemiologists are steeling themselves for the next influenza pandemic, and looking to the past for lessons about control measures. Hence the paper by Christina E. Mills and colleagues elsewhere in this issue (*Nature* **432**, 904–906; 2004), in which the authors have looked at the transmission dynamics of the 1918 pandemic in the United States.

Global outbreaks of influenza occurred three times in the last century. The 1918 pandemic probably killed at least 20 million people: the flu strain responsible spread rapidly, and individuals who became infected were ten times more likely to die than was the case in other pandemics. No wonder people so feared infection — and, for example, that 'influenza camps' were set up, such as that pictured here, at Lawrence, Maine.

The warning signs are there for

the future. Highly pathogenic strains of avian influenza have in very recent times crossed from birds to humans. As virologists Richard J. Webby and Robert G. Webster have put it (*Science* **302**, 1519–1522; 2003): "An old foe has again raised its head, reminding us that our worst nightmare may not be a new one."

How easy would it be to control a new influenza pandemic? By analysing data on deaths from pneumonia and influenza in 45 US cities during 1918, Mills *et al.* estimate the reproductive number (*R*), that is, the number of secondary cases produced by each primary case. This is a measure of viral transmissibility, and given the rapid spread of the 1918 pandemic one might expect it to be rather high. But it turns out that its transmissibility was actually quite modest, with



BETTMANN/CORBIS

each primary case giving rise on average to just three or four secondary cases.

In the absence of global stocks of antiviral drugs and vaccines, warn the authors, control of a similar influenza strain would probably require aggressive measures early

in an outbreak to reduce contacts between persons, whether or not they had already been diagnosed with flu. If a variant virus cropped up with a similar pathogenicity but a much higher transmission rate, we could be in for real trouble.

Rory Howlett

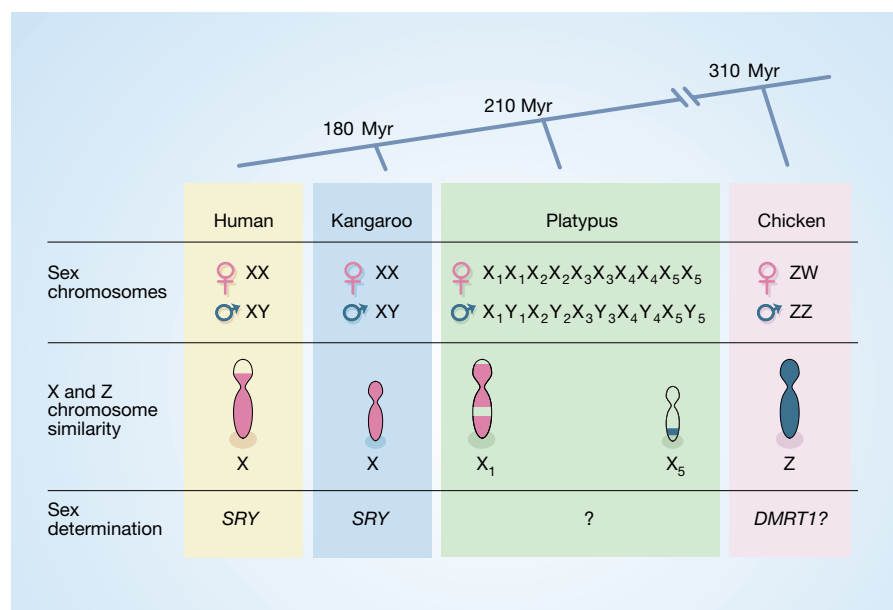


Figure 1 Comparison of sex chromosomes. The four groups represented are birds (the chicken) and examples (human, kangaroo and platypus) of three orders of mammal. The time when each species diverged from their common ancestor is indicated in millions of years (Myr) ago. Pink indicates similarity between X chromosomes; blue, similarity to the Z chromosome. As only one gene from the bird Z chromosome has been mapped to platypus X₅, it is not known whether the platypus chromosomes have other regions that resemble the Z chromosome.

page 914). They discovered ten chromosomes in the chain, five of which are present only in males (Y₁–Y₅). The other five chromosomes in the chain (X₁–X₅) have one copy in males and two in females. These data clearly establish that the platypus' spectacular sex-chromosome composition in females is X₁, X₁, X₂, X₂, X₃, X₃, X₄, X₄, X₅, X₅ and that in males is X₁, Y₁, X₂, Y₂, X₃, Y₃, X₄, Y₄, X₅, Y₅.

During the stage of cell division in which this chromosome chain occurs, like chromosomes form pairs, establish a physical association and then separate, or segregate, to ensure that one of each pair ends up in each daughter cell. In most mammals, although the DNA sequence of the X and Y chromosomes tends to be different, small regions at the ends are identical and allow the two chromosomes to pair. But how do the platypus chromosomes form the intriguing chain structure and properly segregate at meiosis? Grützner *et al.*² found that most of the sex chromosomes have patches of similarity to their immediate neighbours in the chain, suggesting a way that these chromosomes might align with one another to form the meiotic chain.

With so many sex chromosomes, the platypus must have a system for coordinating chromosome separation so that each female-determining sperm ends up with five Xs and each male-determining sperm with five Ys, otherwise random combinations of Xs and Ys would result in unbalanced embryos. The chromosome paints showed that Xs alternate with Ys in the meiotic chain and, importantly, that they separate very efficiently to different sperm² (Fig. 4, page 915).

Such high segregation fidelity of this complicated structure seems remarkable, particularly given that abnormally rearranged chromosomes in other species, in which just two (not five) pairs of chromosomes physically associate at meiosis, frequently result in chromosomes separating unequally or even cause the cells to stop dividing. Perhaps the physical constraints of this large chain make alternative segregation patterns less probable. It is possible that, as hypothesized for a similar structure in evening primrose⁷, the platypus chain forms a 'zigzag' that introduces tension to facilitate segregation of alternate chromosomes. But further studies are required to elucidate the mechanism of separation.

Sex chromosomes (in birds or mammals) are thought to have evolved from a pair of non-sex chromosomes⁹. A series of chromosome rearrangements allowed the X and Y (or Z and W) to evolve independently. The present mammalian Y is a small chromosome with few genes, and the significantly larger X is highly conserved among mammals (Fig. 1), as its evolution has been influenced by X-chromosome inactivation. This is a chromosome-wide gene-silencing process that equalizes gene dosage between males and females.

The platypus meiotic chain probably formed from a succession of additional rearrangements between sex chromosomes and non-sex chromosomes. Models predict that the ancestral sex-chromosome pair should lie at one end of the chain configuration³ — that is, either X₁:Y₁ or X₅:Y₅. The platypus X₁ most closely resembles other

mammalian Xs (ref. 10), but Grützner *et al.*² conclude that the ancestral chromosome pair is actually X₅:Y₅ because, unlike X₁:Y₁, they are not very similar to one another, suggesting that they have had longer to diverge.

Grützner *et al.*² also found that platypus X₅ contains a close relative of the bird candidate sex-determining gene *DMRT1*. Although they acknowledge that *DMRT1* is an unlikely testis-determination gene in platypus, because it is present in two copies in females and only one in males, the authors argue that the significance of this location is in anchoring components of the mammalian X chromosome and bird Z chromosome at either end of the meiotic chain. These data raise the intriguing possibility that, contrary to previous theories⁵, the bird and mammalian sex-chromosome systems are somehow linked. Additional comparative mapping will be necessary to test this hypothesis. The chromosome paints developed here will be a valuable resource for exploring the origins of the chromosomes involved. One might expect that the platypus sex chromosomes will show additional bird Z- and potentially even W-chromosome similarity.

This work raises many questions. First, with the two most obvious candidates, *SRY* and *DMRT1*, eliminated, the search for the platypus sex-determination gene must continue. Its identification could be important for understanding sex determination in other mammals that lack *SRY*¹¹. Another key question is whether the platypus Xs are dosage compensated, because Grützner *et al.* show that only a subset of X_{1–5} sequences are present on Y_{1–5}. Although X₁ seems to be dosage compensated¹, it is unclear whether the mechanism is similar to X-chromosome inactivation in other mammals.

Further studies of the curious meiotic chain in this odd duck of a mammal will provide insights into the evolution of chromosomes, meiosis and sex determination. Who knows what additional links this chain of sex chromosomes will establish?

Laura Carrel is in the Department of Biochemistry and Molecular Biology, Pennsylvania State University College of Medicine, Mail Code H171, Hershey, Pennsylvania 17033, USA.
e-mail: lcarrel@psu.edu

- Grützner, F. & Marshall Graves, J. A. *Curr. Opin. Genet. Dev.* **14**, 642–649 (2004).
- Grützner, F. *et al.* *Nature* **432**, 913–917 (2004).
- Rens, W. *et al.* *Proc. Natl Acad. Sci. USA* **101**, 16257–16261 (2004).
- Koopman, P., Gubbay, J., Vivian, N., Goodfellow, P. & Lovell-Badge, R. *Nature* **351**, 117–121 (1991).
- Fridolfsson, A. K. *et al.* *Proc. Natl Acad. Sci. USA* **95**, 8147–8152 (1998).
- Smith, C. A. & Sinclair, A. H. *BioEssays* **26**, 120–132 (2004).
- Hejnowicz, Z. & Feldman, L. J. *Chromosome Res.* **8**, 165–172 (2000).
- Syren, R. M. & Luykx, P. *Nature* **266**, 167–168 (1977).
- Ohno, S. *Sex Chromosomes and Sex-Linked Genes* (Springer, Berlin, 1967).
- Watson, J. M., Spencer, J. A., Riggs, A. D. & Graves, J. A. *Proc. Natl Acad. Sci. USA* **88**, 11256–11260 (1991).
- Just, W. *et al.* *Nature Genet.* **11**, 117–118 (1995).

Genomics

The silkworm show

Science **306**, 1937–1940 (2004)
<http://silkworm.genomics.org.cn>

The silkworm (*Bombyx mori*) is the latest model organism to have its genome sequence pieced together. The silkworm, second only to the fruitfly *Drosophila* in its importance to insect geneticists, has an estimated 18,510 genes contained within its 28 chromosomes, compared with some 13,379 genes for the fruitfly.

The draft sequence was compiled by a consortium of Chinese researchers, and was no mean feat. Silkworms, which belong to the order Lepidoptera, have a genome sequence totalling almost 430 million base pairs. This is 1.5 times the size of the mosquito sequence and some 3.6 times that of *Drosophila*; both of these are members of the Diptera, which diverged from the Lepidoptera around 300 million years ago.

Of most interest are the genes that govern silk production, a process that has been exploited by mankind for more than 5,000 years. The researchers discovered 1,874 genes that function in the silk glands, only 45 of which were known before in *B. mori*. Many of them seem to encode hormones that regulate the activity of the silk-producing genes themselves. The authors also found 107 genes that seem to have counterparts in that other master spinner, the spider.

Michael Hopkin

Chemistry

Radium singled out

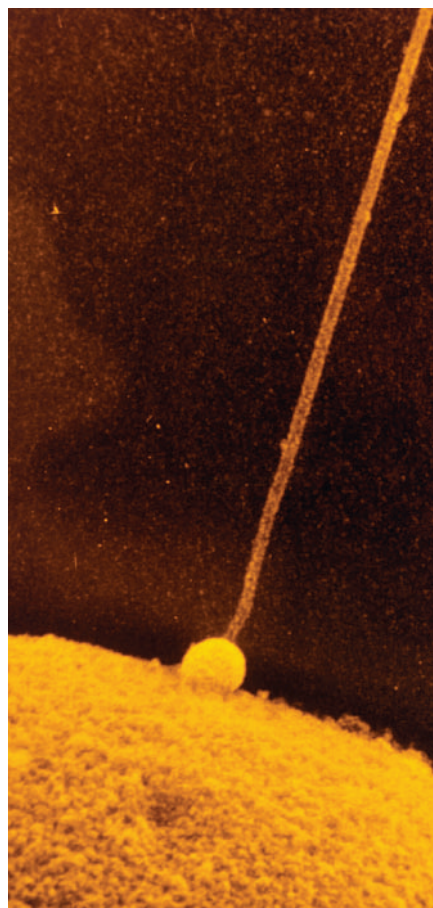
J. Am. Chem. Soc. doi:10.1021/ja0455650 (2004)

Removing radioactive radium from industrial waste-water is tricky — for each radium ion present there may be millions of other metal ions that interfere with the cleaning process.

Fijs W. B. van Leeuwen and colleagues now show that an extraction chemical that builds itself *in situ* can pick radium out of the crowd. The chemical is an isoguanine derivative, which self-assembles into layered complexes and uses the radium ions as a template.

The complexes trap radium over a much wider pH range than has been possible with conventional extraction methods, and have a higher selectivity. For example, in a mixture containing a million times more magnesium than radium, van Leeuwen *et al.* could still remove all of the radium. Barium, which is close in size to radium, is an even tougher proposition. But the isoguanine complex still extracted almost all the radium in a mixture, even when the radium was outnumbered by 10,000 to one.

Mark Peplow



Developmental biology

Peroxide producer

Dev. Cell **7**, 801–814 (2004)

The fertilized egg must keep a troop of invading sperm at bay. To do this, it undergoes a series of chemical reactions that hardens the outside of the cell into a protective shield. In many animals this is achieved by using hydrogen peroxide. For decades, scientists have searched for clues to how the peroxide is produced and what prevents it from damaging the fertilized egg. An investigation of sea urchins by Julian L. Wong *et al.* now sheds light on this reproductive riddle.

The authors scanned the sea urchin's DNA for genes encoding enzymes that could catalyse these types of peroxide reactions. An enzyme that they dubbed Udx1, for 'urchin dual oxidase 1', caught their attention. They showed that Udx1 is activated after fertilization and starts to generate hydrogen peroxide. This peroxide, in turn, is used to crosslink proteins to create a tough layer surrounding the egg.

So what makes Udx1 a 'dual' enzyme? Although it can convert oxygen into peroxide, it can also, using different domains of the enzyme, cleave this toxic compound into harmless water molecules, protecting the embryo.

Roxanne Khamis

Photovoltaics

Solar cell, heal thyself

Appl. Phys. Lett. **85**, 5218–5230 (2004)

If money is no object, photovoltaic cells can be impressively efficient at turning sunlight into electricity. A multi-decker cell made from indium gallium phosphide, indium gallium arsenide and germanium can attain efficiencies of more than 29%, well above the 16% efficiencies typical of commercial silicon cells. The cost is far greater too — but if the cell is powering instrumentation on a space satellite it's a small price to pay. Such a cell can perform even better if a layer of AlInGaP is added, which captures the high-energy photons in the solar spectrum.

There's another reason why silicon solar cells don't do well in space, however: they are degraded by the harsh radiation they encounter. InGaP and InP have been shown to be relatively resistant to such radiation, and now Aurangzeb Khan *et al.* show that AlInGaP shares this important property, making such high-efficiency, multilayered solar cells feasible for space applications.

The main defects produced in AlInGaP by an electron beam, which mimics the radiation environment of space, are so-called hole traps, which can act as sites for wasteful recombination of light-induced electrons and holes in the semiconductor. But these defects disappear in an annealing process, making the material more or less self-healing.

Philip Ball

Molecular biology

Sun protection factor

Proc. Natl Acad. Sci. USA
 doi:10.1073/pnas.0406304101 (2004)

People who suffer from a genetic disease known as xeroderma pigmentosum have a 1,000-fold increased risk of developing cancer in sun-exposed areas, particularly the skin. This is because their cells lack a functional copy of a crucial enzyme that is involved in repairing DNA damage induced by ultraviolet light. The lifespan of those with this condition is reduced by 30–40 years.

Although a cure in humans remains elusive, Maria Carolina N. Marchetto and her colleagues have just seen success in treating mice engineered to bear this genetic defect. The authors injected a modified adenovirus carrying the human XPA gene, one of the genes known to be affected in xeroderma pigmentosum, into the skin of such mice. They found that the injection increased the expression of the XPA protein — and helped to protect the mice against skin tumours after exposure to UVB light.

Roxanne Khamis

Cage enrichment and mouse behaviour

Test responses by laboratory mice are unperturbed by more entertaining housing.

Mice housed in standard cages show impaired brain development, abnormal repetitive behaviours (stereotypies) and an anxious behavioural profile, all of which can be lessened by making the cage environment more stimulating^{1–3}. But concerns have been raised that enriched housing might disrupt standardization and so affect the precision and reproducibility of behavioural-test results (for example, see ref. 4). Here we show that environmental enrichment increases neither individual variability in behavioural tests nor the risk of obtaining conflicting data in replicate studies. Our findings indicate that the housing conditions of laboratory mice can be markedly improved without affecting the standardization of results.

We raised female mice of two inbred strains (C57BL/6J and DBA/2) and their F₁-hybrids (B6D2F1) in three different laboratories in small, standard cages or large, enriched cages. When the mice reached adulthood, we subjected them to four of the most common behavioural tests used in drug screening and in behavioural phenotyping of transgenic and 'knockout' mice: elevated O-maze, open-field test, novel-object test and place navigation in the water maze (see supplementary information for methods). Each laboratory independently ordered three batches of 48 mice (8 mice per strain and housing condition) to run three independent replicates for each test (total number of mice was 432).

To test the effects of enriched housing on the detection and reproducibility of genetic differences in behaviour between the mice, we split the data by housing condition and calculated the proportion of variance for each replicate in behavioural measures contributed by within-group variability and by laboratory $\times$ strain interactions. Proportions of variance for representative measures of the four tests were then compared between enriched and standard housing conditions.

Figure 1 shows a summary of the results. Within-group variability contributed between 40% and 84% (average 60%) to total variance. With an average of 7.6%, the contribution of strain $\times$ laboratory interactions was considerably smaller and also less variable. However, within-group variability was unaffected by enriched housing (except for faecal bolus counts on the O-maze). This indicates that enrichment did not decrease the sensitivity of tests for detecting genetic differences and also shows that bare cages fail to remove individual differences in behaviour. Enrichment had no significant effect on the proportions of variance contributed by strain $\times$

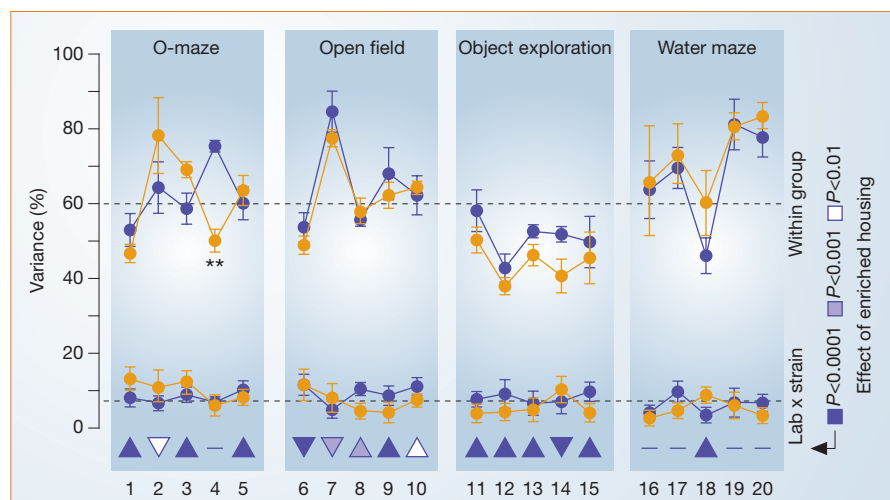


Figure 1 Mean ($\pm$ s.e.m.) proportion of variance in representative measures of four behavioural tests, contributed by within-group variability and laboratory $\times$ strain interactions, in a multilaboratory study comparing female DBA/2, C57BL/6J and B6D2F1 mice housed under standard (orange) and enriched (blue) conditions. To compare the partitioning of variance under standard and enriched conditions, data were split by housing condition, and replicates were treated as independent observations. Variance proportions were arcsine-square-root transformed and compared using *t*-tests. In the object-exploration test, within-group variability of standard housed mice was decreased by a 'floor' effect (data near observable minimum) because many mice showed object avoidance. Triangles indicate direction and significance of enrichment effects on each variable. Double asterisks, significant difference in individual variability. Representative measures: 1, total head dips; 2, percentage of protected head dips; 3, percentage of open arm entries; 4, faecal-bolus count; 5, total path travelled; 6, centre avoidance, first 10 min; 7, centre avoidance, habituation; 8, path travelled, first 10 min; 9, path travelled, habituation; 10, percentage of time spent running/walking; 11, horizontal-object exploration; 12, vertical-object exploration; 13, percentage of path in object zone; 14, object-exploration distance; 15, corner distance; 16, swim-path length; 17, percentage of time spent near wall; 18, average swim speed; 19, probe: annulus crossings; 20, probe: target proximity. Animals were checked daily and remained healthy throughout the experiment. See supplementary information for further details.

laboratory interactions, and the direction of differences varied across measures, indicating that enrichment did not increase the risk of obtaining conflicting results between labs.

Strain and housing effects on test performance were analysed using a four-way factorial analysis-of-variance model with strain, housing conditions, laboratory and replicate as between-subject factors. As in an earlier multilaboratory study⁵, we found significant strain $\times$ laboratory interactions on many variables. However, these were quantitative rather than qualitative in nature, reflecting differences in effect magnitude rather than direction of the effects (see supplementary information). Likewise, enrichment effects were highly consistent across laboratories (Fig. 1, and see supplementary information).

Between-laboratory effects (contributing on average 5.2%) and replicate effects (3.1%) made comparably small contributions to total variance, indicating that standardization between labs was nearly as good as standardization within labs. This was surprising because only cage type, enrichment protocol, light phase, test system and test protocol were equated across labs. This result casts doubt on the effectiveness of excessive environmental harmonization to improve between-laboratory comparability of behavioural data^{6,7}

and, together with the large individual variability, underscores the need for sufficiently large sample sizes if subtle genetic effects (as often sought in 'knockout' and transgenic mice) are to be detected reliably.

Our findings argue against concerns that environmental enrichment might disrupt standardization, which may have hindered the implementation of enriched housing, despite its known advantages to animals^{1–3}. Our findings should be generally applicable, for example in drug-screening or lesion studies. And they should also apply to animals' physiology or anatomy, which are, in any case, less sensitive than behaviour to environmental perturbations.

It remains to be seen whether our conclusions also apply to male mice, who may respond to enrichment with enhanced dominance behaviour and aggression⁸. But for females, environmental enrichment should improve the animals' well-being without reducing the precision and reproducibility of the data derived from them, while attenuating abnormal brain function² and anxiety³ — both of which are potential confounds in animal experiments.

David P. Wolfer*, Oxana Litvin†, Samuel Morf*, Roger M. Nitsch†, Hans-Peter Lipp*, Hanno Würbel‡

*Division of Neuroanatomy and Behaviour,
Institute of Anatomy, University of Zurich,
8057 Zurich, Switzerland

†Division of Psychiatry Research, University of
Zurich, 8008 Zurich, Switzerland

‡Animal Welfare and Ethology Group, Institute of
Veterinary Physiology, University of Giessen,
35392 Giessen, Germany

e-mail: hanno.wuerbel@vetmed.uni-giessen.de

1. Van Praag, H., Kempermann, G. & Gage, F. H. *Nature Rev. Neurosci.* **1**, 191–198 (2000).

2. Würbel, H. *Trends Neurosci.* **24**, 207–211 (2001).

3. Chapillon, P., Manneche, C., Belzung, C. & Caston, J. *Behav. Genet.* **29**, 41–46 (1999).

4. Gärtner, K. in *Proc. Int. Joint Meeting 12th ICLAS General Assembly and Conference & 7th FELASA Symposium, SECAL, Madrid 207–210* (1999).

5. Crabbe, J. C., Wahlsten, D. & Dudek, B. C. *Science* **284**, 1670–1672 (1999).

6. van der Staay, F. J. & Steckler, T. *Genes Brain Behav.* **1**, 9–13 (2002).

7. Würbel, H. *Genes Brain Behav.* **1**, 3–8 (2002).

8. Van Loo, P. L. P., Van Zutphen, L. F. M. & Baumans, V. *Lab. Anim.* **37**, 300–313 (2003).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Neurosurgery

Functional regeneration after laser axotomy

Understanding how nerves regenerate is an important step towards developing treatments for human neurological disease¹, but investigation has so far been limited to complex organisms (mouse and zebrafish²) in the absence of precision techniques for severing axons (axotomy). Here we use femtosecond laser surgery for axotomy in the roundworm *Caenorhabditis elegans* and show that these axons functionally regenerate after the operation. Application of this precise surgical technique should enable nerve regeneration to be studied *in vivo* in its most evolutionarily simple form.

Femtosecond laser pulses provide high

peak intensities that reduce the energy threshold for tissue removal (ablation)³ and enable laser surgery to be carried out with a low-energy source^{4,5}. (For methods, see supplementary information.) We successfully cut single axons inside *C. elegans* by using pulse energies of 10–40 nanojoules at the specimen and tightly focused, 200-femtosecond, near-infrared laser pulses. This results in the vaporization of axon volumes of about 0.1–0.3 femtolitres, assuming an average axon diameter of 0.3 micrometres (see supplementary information). Dye-filling of axotomized neurons confirmed that the observed axon gaps are not due to photobleaching, but to physical disconnection of the axons (see supplementary information).

The minimum energy used is consistent with measured optical breakdown thresholds in transparent materials^{3,6}. At these low energies, we would expect mechanical effects due to plasma expansion and shock waves to be significantly reduced^{5,6} with respect to other laser-surgery techniques that require much higher energies (for example, 0.4 microjoules with 0.5-nanosecond pulses⁷). The use of pulses at a low repetition rate (1 kilohertz, 10 microwatts average power) should reduce heat accumulation and extended thermal damage to the environment. We were able to cut individual processes within a few micrometres of each other without damaging the nearby processes (see supplementary information).

The D-type motor neurons in L4 larval-stage worms were selected as targets for laser surgery. These neurons have ventral cell bodies and extend circumferential axons towards the dorsal side; they form synapses to body muscles⁸. We cut the circumferential axons (labelled with green fluorescent protein⁹) at their mid-body positions, leaving the rest of the axon intact (Fig. 1a). Both ends of the severed axons initially retracted. Among 52 operated axons (in 11 worms),

54% regrew towards their distal ends within 12–24 hours (Fig. 1b). Axons that showed partial, aberrant, or no regrowth within 24 hours did not show further improvement over longer observation times (up to 36 hours).

To evaluate functional recovery associated with nerve regeneration, we tested the behaviour of operated worms as it related to motor-neuron function. Loss of D-neuron function results in simultaneous contraction of dorsal and ventral body muscles ('shrinker' phenotype)¹⁰, which prevents backward locomotion. Operated worms (17 in total) showed this expected 'shrinker' phenotype immediately after axotomy (15 axons per worm), whereas sham-operated animals (6 in total) moved like wild-type worms.

Remarkably, the locomotion of operated worms improved, approaching that of the wild type within 24 hours of surgery (Fig. 1c), indicating that the regenerated axons were functional (see movie in supplementary information). By contrast, the shrinker phenotype caused by laser ablation of D-neuron cell bodies did not recover after 48 hours (results not shown). The correlation of axonal regrowth with behavioural recovery in *C. elegans* indicates that these nerves must have regenerated.

Femtosecond laser axotomy is a new technique that can be performed with 100% efficiency, sub-micrometre precision and high speed. As simple organisms such as *C. elegans* have amenable genetics, application of the femtosecond laser axotomy technique we describe here should help in the rapid identification of genes and molecules that affect nerve regeneration and development.

Mehmet Fatih Yanik*, Hulusi Cinar†, Hediye Nese Cinar†, Andrew D. Chisholm†, Yishi Jin†‡, Adela Ben-Yakar*§

*Department of Applied Physics, Ginzton Laboratory, Stanford University, Stanford, California 94305, USA

†Department of MCD Biology, Sinsheimer Laboratories, University of California, Santa Cruz, California 95064, USA

‡Howard Hughes Medical Institute, University of California, Santa Cruz, California 95064, USA

§Department of Mechanical Engineering, University of Texas at Austin, Austin, Texas 78712, USA

e-mail: ben-yakar@mail.utexas.edu

1. Horner, P. J. & Gage, F. H. *Nature* **407**, 963–970 (2000).

2. Bhatt, D. H., Otto, S. J., Depoister, B. & Fetcho, J. R. *Science* **305**, 254–258 (2004).

3. Perry, M. D. et al. *J. Appl. Phys.* **85**, 6803–6810 (1999).

4. Tirlapur, U. K. & König, K. *Nature* **418**, 290–291 (2002).

5. Shen, N. Thesis, Harvard Univ. (2003).

6. Vogel, A. et al. *Appl. Phys. B* **68**, 271–280 (1999).

7. Colombelli, J., Grill, S. W. & Stelzer, E. H. K. *Rev. Sci. Instrum.* **75**, 472–478 (2004).

8. White, J. G., Southgate, E., Thomson, J. N. & Brenner, S. *Phil. Trans. R. Soc. Lond. B* **314**, 1–340 (1986).

9. Huang, X., Cheng, H. J., Tessier-Lavigne, M. & Jin, Y. *Neuron* **34**, 563–576 (2002).

10. McIntire, S. L., Jorgensen, E., Kaplan, J. & Horvitz, H. R. *Nature* **364**, 337–341 (1993).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

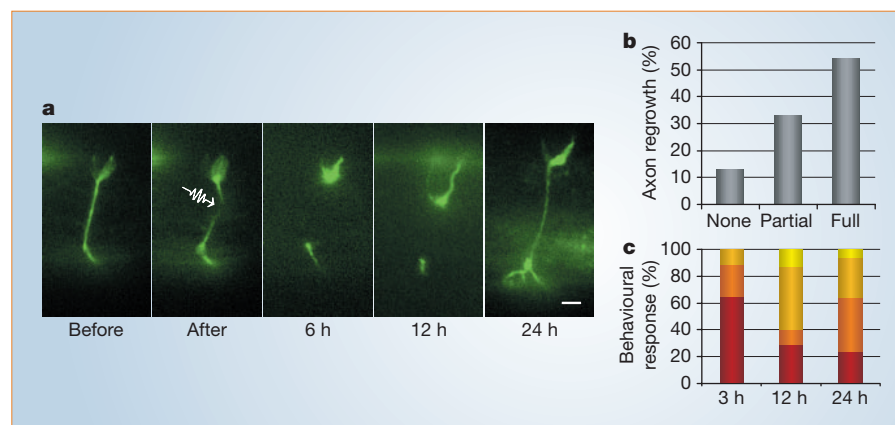
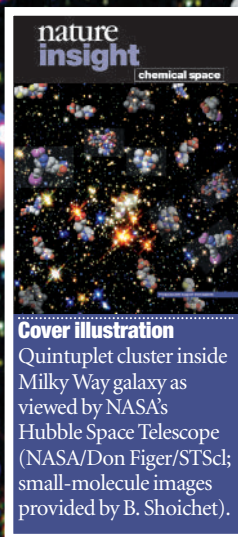


Figure 1 Femtosecond laser axotomy in *Caenorhabditis elegans* worms using 100 pulses of low energy (40 nanojoules) and short duration (200 femtoseconds) and a repetition rate of 1 kilohertz. **a**, Fluorescence images of axons labelled with green fluorescent protein before, immediately after, and in the hours following axotomy. Arrow indicates point of severance. Scale bar, 5 μ m. **b**, Statistics of axon growth 24 h after axotomy, based on fluorescence images ($n = 52$ axons). **c**, Time-course analysis of backward motion of worms following axotomy. Seventeen worms were scored blindly at different time points (for criteria, see supplementary information). Improvement in backward motion was graded as four levels from 'shrinker' behaviour (dark red) up to 'wild type' behaviour (yellow) in the hours following axotomy.

nature insight

chemical space



“Space”, as Douglas Adams famously said “is big. You just won't believe how vastly, hugely, mind-bogglingly big it is”. Change ‘space’ to ‘chemical space’, and his statement has similar resonance: the total number of possible small organic molecules that populate ‘chemical space’ has been estimated to exceed 10^{60} — an amount so vast when compared to the number of such molecules we have made, or indeed could ever hope to make, that it might as well be infinite. So, it is not surprising that our exploration of chemical space has so far been extremely limited.

Taking the analogy further, just as much of astronomical space is a void, much of chemical space contains nothing of biological interest. But rarely, and often through serendipity rather than design, we have identified ‘stars’ in chemical space — molecules that can modulate biological processes. These molecules have formed much of the basis of our fight against disease and have greatly aided our understanding of biological systems.

But such successful finds have been hard to come by, in part because of our lack of understanding of chemical space. Given that its enormous size makes a thorough exploration of chemical space impossible, a key question is how we should best direct our efforts towards regions of chemical space that are most likely to contain molecules with useful biological activity. This question is a central theme of the articles in this Insight, which were inspired by the Horizon Symposium on ‘Charting Chemical Space: Finding New Tools to Explore Biology’, the fourth in a series of unique scientific discussion meetings run by Nature Publishing Group and Aventis.

We are pleased to acknowledge the support of Aventis in producing this Insight. As always, *Nature* carries sole responsibility for all editorial content and peer review.

Peter Kirkpatrick Chief Editor, *Nature Reviews Drug Discovery*

Clare Ellis Business Editor, *Nature Reviews Drug Discovery*

- 824 introduction**
Chemical space and biology
C. M. Dobson
- 829 review articles**
Lessons from natural molecules
J. Clardy & C. Walsh
- 838 Natural and engineered nucleic acids as tools to explore biology**
R. R. Breaker
- 846 Exploring biology with small organic molecules**
B. R. Stockwell
- 855 Navigating chemical space for biology and medicine**
C. Lipinski & A. Hopkins
- 862 commentary**
Virtual screening of chemical libraries
B. K. Shoichet

Editor, *Nature*: Philip Campbell
Insights Publisher: Sarah Greaves
Insights Editor: Lesley Anson
Consultant Editor: Joshua Finkelstein

Production Editor: Lucy Odling-Smee
Art Editor: Martin Harrison
Layouts: Nadia Rooney
Diagrams: Nik Spencer

Sponsorship: Claire Hines
Production: Sue Gray
Marketing: Claire Aspinall
Editorial Assistant: Timothy Gibbs

Chemical space and biology

Christopher M. Dobson

Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK (e-mail: cmd44@cam.ac.uk)

Chemical space — which encompasses all possible small organic molecules, including those present in biological systems — is vast. So vast, in fact, that so far only a tiny fraction of it has been explored. Nevertheless, these explorations have greatly enhanced our understanding of biology, and have led to the development of many of today's drugs. The discovery of new bioactive molecules, facilitated by a deeper understanding of the nature of the regions of chemical space that are relevant to biology, will advance our knowledge of biological processes and lead to new strategies to treat disease.

Living systems have evolved over several billion years to carry out carefully controlled chemistry in an aqueous environment at temperatures almost exclusively between zero and 100 °C. Under these conditions and unaided, many of the chemical reactions that are essential to life would not occur at perceptible rates, and most would not result in specific and reproducible products. Enzymes, along with other proteins and some nucleic acids, are used by natural biological systems to achieve this control; these macromolecules are responsible for the synthesis, transport and degradation of virtually every chemical compound in the biological environment¹.

However, the chemical compounds used by biological systems represent a staggeringly small fraction of the total possible number of small carbon-based compounds with molecular masses in the same range as those of living systems (that is, less than about 500 daltons). Some estimates of this number are in excess of 10^{60} (ref. 2). The simplest living organisms can function with just a few hundred different types of such molecule, and fewer than 100 account for nearly the entire molecular pool^{3,4}. Moreover, it seems that the total number of different small molecules within our own bodies could be just a few thousand⁴. So, it is clear that, at least in terms of numbers of compounds, 'biologically relevant chemical space' is only a minute fraction of complete 'chemical space' (see Box 1 for a definition of the terms used in this Insight). It is remarkable that so many complex processes can be carried

out with such a limited number of molecules, and that biological chemistry can be so rich and diverse despite the relatively limited range of reactions that seem to have been exploited during the evolution of living systems (see Box 2 for a discussion of why particular types of chemistry might have emerged as the basis of life).

Similarly, as revealed by the recent triumphs of a variety of international sequencing projects, the genomes of the simplest living systems encode the sequences of less than 1,000 different proteins and the human genome about 100 times more⁵ — numbers that are minute when compared with the total number of proteins that could theoretically exist. As there are 20 different types of amino acid and the average size of a natural protein is about 300 residues, this number is a staggering 20^{300} or more than 10^{390} , and if only a single molecule of each of these polypeptides were to be produced, their combined mass would vastly exceed that of the known universe. Natural proteins are therefore also a very select group of molecules.

The characteristics of this select group of natural proteins are linked to those of the small molecules that are used in living systems, and to those of the relatively small number of synthetic small molecules that we have developed into drugs. Understanding this link will help us answer the question of how we can best use the powerful new methods that are emerging to probe biological systems, both to understand the fundamental processes of life and to develop new strategies to treat disease.

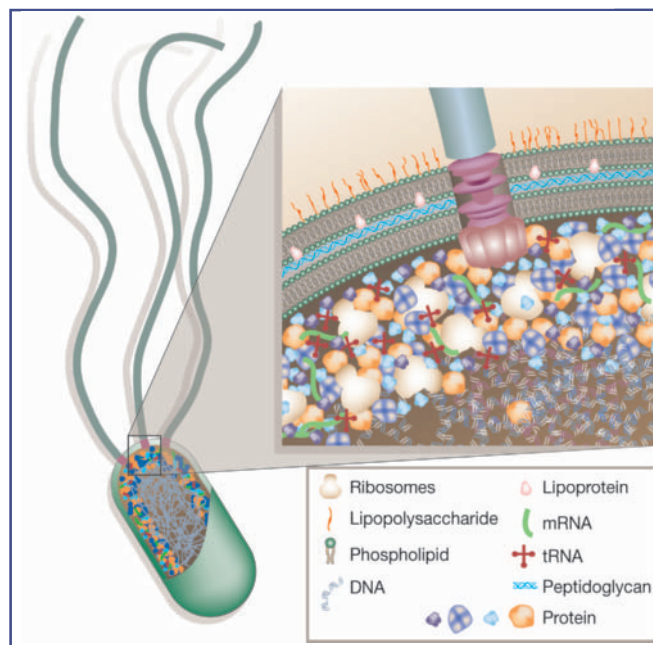


Figure 1 Schematic representation of a crowded cell. An array of different molecules can function independently under extremely crowded conditions, partly because of judicious distributions of oppositely charged polar groups on the molecular surfaces³⁸. However, such systems are in some ways extremely fragile. For example, a mutation that alters just one amino acid in the haemoglobin molecule (replacing a charged carboxylic acid with a methyl group) can stimulate massive aggregation and give rise to a fatal genetic disease, sickle-cell anaemia^{8,39}. More generally, many disorders of old age, most famously Alzheimer's disease, result from the increasingly facile conversion of normally soluble proteins into intractable deposits that occur particularly as we get older (see <http://www.horizonsymposia.com/> for the Horizon Symposium 'Protein Folding and Disease', and ref. 40). Many of these aggregation processes involve the reversion of the unique biologically active forms of polypeptide chains into a generic and non-functional 'chemical' form⁴¹. Adapted with permission from D. Goodsell.

Box 1

Glossary of important terms relevant to chemical space and biology
Bioavailability

The fraction or percentage of an administered drug or other substance that becomes available to the target tissue after administration.

Biologically relevant chemical space

Those parts of chemical space in which biologically active compounds reside.

Chemical genetics

The study of gene-product function in a cellular or organismal context using a set of exogenous ligands, often known as chemical tools.

Chemical library

A collection of chemical compounds.

Chemical space

Chemicals can be characterized by a wide range of 'descriptors', such as their molecular mass, lipophilicity (their affinity for a lipid environment) and topological features. 'Chemical space' is a term often used in place of 'multi-dimensional descriptor space': it is a region defined by a particular choice of descriptors and the limits placed on them. In the context of this Insight, chemical space is defined as the total descriptor space that encompasses all the small carbon-based molecules that could in principle be created.

Combinatorial chemistry

The generation of large collections or 'libraries' of compounds by combinations of a set of smaller chemical structures, known as 'building blocks'.

Druggability/druggable target

The feasibility with which a macromolecular target can be modulated by a small molecule that has appropriate properties to be developed into a drug.

Drug-like

Sharing certain characteristics with other molecules that act as drugs. The exact set of characteristics — size, shape and solubility in water and organic solvents — varies depending on who is evaluating the molecules.

Genome

All the genetic material in the chromosomes of a particular organism.

High-throughput screening

In high-throughput screening, large libraries of chemical compounds (typically 10,000 to 100,000) are screened in a biological assay, for example, for their ability to bind to a particular protein or to inhibit a particular cellular process.

Hit

An active compound that exceeds a certain threshold value in a given assay; for example, more than 90% inhibition of an enzyme's activity.

Lead

A chemical structure or series of structures that demonstrate activity and selectivity in a biological screen. In drug discovery, a lead is used as a basis for chemical optimization, with the aim of identifying a clinical candidate.

Lipinski's rules

Lipinski's analysis of the World Drug Index led to the 'rule of five'¹⁵. This identifies several key properties that should be considered for small molecules that are intended to be orally administered. These properties are: molecular mass less than 500 daltons; number of hydrogen-bond donors less than 5; number of hydrogen-bond acceptors less than 10; calculated octanol/water partition coefficient (an indication of the ability of a molecule to cross biological membranes) less than 5.

Natural product

A chemical substance produced by a living organism. This term is often used in reference to small chemical substances found in nature that have distinct pharmacological effects, such as the antibiotic penicillin.

Proteome

The complete set of proteins that can be expressed by the genetic material of an organism.

RNA interference (RNAi)

A process by which double-stranded RNA silences specifically the expression of homologous genes.

Chemistry in a biological environment

A crucial factor in understanding the nature of living systems is that biological molecules do not act in isolation in the dilute solutions familiar to most chemists. Instead, they are packed together to an extraordinary degree within cells^{6,7}. Indeed, the concentration of macromolecules inside cells can amount to several hundred grams per litre. Many of us may have been astonished during our school days to learn that our bodies are more than 70% water, but how many of us wondered at the difficulty of making a 30% solution of molecules that are rich in hydrocarbon derivatives and other hydrophobic groups? A space-filling representation of a typical cell (Fig. 1) illustrates how molecular species are crowded together in its complex organizational structure^{8,9}. Such 'molecular crowding' is likely to be important in many facets of biological chemistry. For example, binding affinities and the rates of self-assembly can change by orders of magnitude as a result of this phenomenon. Crowding is therefore an important factor to consider when using data derived from *in vitro* studies in dilute solution to understand processes taking place *in vivo*^{6,7}. Moreover, biological systems are increasingly being considered as highly interconnected sets of interactions (as shown, for example, by the emergence of 'systems biology') in contrast to the reductionist view of much of traditional biochemistry¹⁰. In addition, considerable efforts are being

made to understand the astonishing ability of biological molecules to self-assemble and generate functional entities ranging from folded proteins to whole organisms¹¹.

Techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) and mass spectrometry have already revolutionized our understanding of the structure and function of biological molecules. It is now becoming possible to examine the ultrastructure of cells in remarkable detail, primarily through the development of modern imaging techniques¹². Of particular importance are methods based on fluorescence emission. These can be used together with confocal microscopy to identify and track an increasingly wide range of molecules (both large and small) within their biological environments. Perhaps the most dramatic technique, however, is that based on electron microscopy: 'cryoelectron tomography' is now beginning to allow us to visualize, within a cell, molecular assemblies such as actin, which provides cells with their internal structures, and ribosomes, the complexes of proteins and nucleic acids that are responsible for all protein synthesis¹³. Along with these experimental approaches, computational procedures are being developed to simulate the behaviour of molecules within whole cells or indeed whole organisms¹⁴. Further developments of this type will undoubtedly lead to a deeper understanding of how cellular components of all types interact with each other. Even without

such information, however, the high density of molecules in cells is a remarkable phenomenon that must be borne in mind when we attempt to perturb their behaviour for therapeutic purposes.

The challenges of drug discovery

Although some therapeutic agents are designed to increase the natural concentrations of key biological molecules that are depleted in particular disease states (for example, insulin), the primary objective of most pharmaceutical chemistry is to generate new compounds that can modulate disease processes. Most prized are relatively small molecules (only a small percentage of orally administered drugs have molecular masses above 500 daltons¹⁵) whose properties enable them to interact with and perturb the function of given biological molecules. It is equally important, however, that these compounds do not interact with most other molecules and generate potentially adverse side effects. The immensity of this task is illustrated by the schematic illustration in Fig. 1.

Box 2

The chemistry of life

One of the most fundamental questions relating to biological diversity is why particular types of molecule have emerged as those on which the chemistry of all life forms is based. It is clear that solubility in water is a key issue. Although 99% of the atoms within a biological system are C, H, O or N, more than 20 other elements are essential to life. All these elements are (or were when life on Earth began) relatively abundant in the Earth's crust, the sea or the atmosphere, and their ions or common compounds are soluble in water⁴². Solubility in water is also likely to be a major reason why many of the small organic molecules used by biological systems (including the amino acids) are derivatives of simple carboxylic acids and organic amines; these groups are normally charged, and therefore hydrophilic, at physiological pH. Similarly, many others are charged derivatives of phosphoric acid⁴³, the chemical entity that is also the precursor of ATP, the chief energy store in biology, and the scaffold for DNA and RNA. The unique properties of water also cause other derivatives of phosphoric acid, the phospholipids, to assemble into bilayers that are the key components of all biological membranes. The energetic advantage of burying hydrophobic groups away from water in the interior of a closely packed structure is also an important driving force in protein folding^{1,41}. To allow folding, a significant proportion of the 20 amino-acid side chains incorporated into natural proteins are very hydrophobic, and the rest, many of which end up on the surface of folded proteins, are to varying degrees hydrophilic.

The chemical properties of the various side chains of proteins, along with a selection of metal ions and cofactors that can be incorporated into the folded structures, not only permit folding but also define the fundamental chemistry of life. The side chains of the natural amino acids, which are the same in every living organism, contain only a small selection of the functional groups that are familiar from any chemistry textbook: a methyl (but not an ethyl) group; an isopropyl (but not an *n*-propyl) group; a primary and a secondary alcohol; a thiol and an imidazole group; two carboxylic acids and so on³⁹. But why this particular set of 20 chemical groups? Do these groups have the unique range of properties required to catalyse all the reactions needed for life to occur? Or did they arise by chance and has life on Earth been too short to allow the evolution of a wider range of chemical entities? The answers to such questions have long been the subject of speculation, but are now beginning to be probed directly by experiment. One remarkable new approach exploits the usual mechanism of protein synthesis in bacteria to generate proteins containing new types of amino acid⁴⁴. It will be fascinating to learn what additional chemical tasks such organisms can perform, and how they respond to selective pressure in laboratory experiments that simulate natural evolution. Undoubtedly, such forays into 'abnormal' biology will shed light on 'normal' biological evolution and function, and indeed on the types of novel chemical entity that can interact selectively with natural biomolecules.

The natural products of different organisms — largely plants and bacteria — or their derivatives have been the staple tools of healers from the dawn of history until the birth of modern synthetic chemistry in the nineteenth century. Now, with the immense developments in combinatorial methods over the past decade or so, huge arrays of new molecules can be produced in relatively short periods of time^{16,17}. Together with rapid screening methods, the drug-discovery process has been moving into uncharted territory; seemingly endless numbers of potentially active compounds are becoming available. As our knowledge of even the most complex aspects of biology at a molecular level expands, we can increasingly use rational arguments in the design of potential therapies and of new molecules that are promising to test or screen¹⁸. Despite such expert knowledge, the scale of the procedures needed to find appropriate compounds is remarkable; some individual drug companies screen millions of potential compounds each year against a range of targets, and even then, success is not guaranteed. As we have seen, however, such numbers are insignificant compared with the total number of possible small organic molecules. In addition, even the biggest libraries of compounds used in screening may not reflect the rich chemical diversity of the much smaller numbers of natural products¹⁹ (Fig. 2). It is clear, therefore, that reliable computational approaches to sift through much larger numbers of more varied compounds would be of tremendous value in drug discovery. Once likely candidates for a given purpose are identified, experimental screening procedures could then be focused on a much smaller range of selected compounds. As Shoichet discusses in a commentary in this issue (page 862), the examination of molecules *in silico* for their ability to bind to specific targets already plays an important part in screening strategies, although such 'virtual screening' approaches have yet to achieve their full potential in the drug-discovery process.

Despite the many advances in technology, the cost of generating new drugs is inexorably rising, leading to ever greater pressure on pharmaceutical companies to focus on developing therapies primarily for the common diseases of wealthy countries^{20,21}. Those suffering from rare diseases, and indeed the vast number of people in poorer countries, particularly in the tropics, are all too often neglected in the continuing fight against infection and disease. But despite the evidence that the new techniques entering the pharmaceutical industry have not yet been a panacea for the drug-discovery process²², it is still early days. We have yet, for example, to reap the real benefits of the recent revolutions in genomics and proteomics, which promise to identify a much greater number of well-characterized molecular targets for therapeutic intervention²³. Indeed, the number of new targets that have emerged in recent years within the pharmaceutical industry as a whole is remarkably small. For example, between 1994 and 2001, just 22 drugs that modulate new targets were approved²⁴. So far, analyses have revealed that the total number of human proteins against which drugs have been targeted is less than 500 (ref. 25), a small percentage of the estimated total number of proteins in the human body. Although expert opinions differ as to the total number of possible 'druggable' targets, it is certainly larger than the number currently known^{25,26}.

Chemical 'tools' for biological systems

One of the potential problems with the new types of organic compound that are now being explored as drugs is that they may be extremely potent when tested against isolated targets in the laboratory environment, but within the complex cellular milieu (Fig. 1), they might interact with cellular components other than the desired target. The small molecules found naturally in biological systems, often called 'natural products', have at least been through the evolutionary mill and are perhaps less likely to interact in a damaging manner with common components of living systems, such as membranes or DNA. Indeed, of all drugs licensed over the past 20 years, around 30% are natural products or natural-product derivatives. If we include compounds 'inspired by' natural products, the fraction rises to almost twice this number²⁷ (see also the review in this

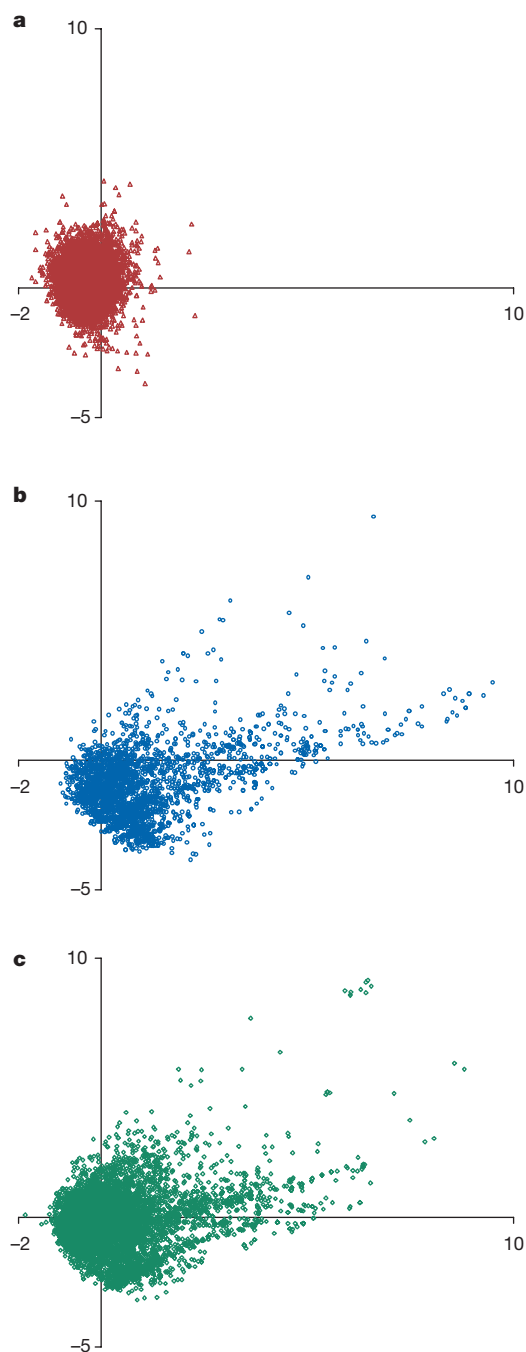


Figure 2 Comparison of the properties of different classes of molecule. A large database that contained compounds from combinatorial chemistry (**a**), natural products (**b**) and drugs (**c**) was analysed on the basis of a variety of molecular properties¹⁹. To visualize the diversity of these compounds on the basis of these properties, a statistical approach known as principal component analysis was used. Plots of the first two principal components — which explain about 54% of the variance in the properties analysed — are shown. Combinatorial compounds cover a well-defined region in diversity space given by these principal components. Both drugs and natural products cover all this space, as well as a much larger additional region of space. It is of particular interest to note the similarity of the plots of natural products and successful drug molecules. Adapted with permission from ref. 19.

issue by Clardy and Walsh, page 829). Interestingly, a comparison of the properties of drugs, natural products and combinatorial chemistry libraries shows that combinatorial compounds typically cover a significantly smaller area of chemical space than either drugs or natural products¹⁹ (Fig. 2). This suggests that by aiming to mimic some properties of natural compounds, new combinatorial compounds could be made that are substantially more diverse and that have greater biological relevance¹⁹ than those currently known.

Remarkably, however, it has been estimated that only 0.1% of all bacterial strains — the richest source of new biological molecules — has been cultured and analysed²⁸. Thus, as Clardy and Walsh discuss in this issue (page 829), there is a vast harvest of new natural products, perhaps running to millions of new compounds, waiting to be gathered from previously unexplored strains of living organisms (mainly bacteria, plants and fungi). Moreover, there are now opportunities to manipulate nature's 'production lines', for example, by using mutagenesis and gene shuffling to induce microorganisms to create new biologically active molecules, and hence to generate large libraries of new 'natural products'.

One of the most important aspects of the development of new techniques and technologies is that they can be used for two distinct but highly complementary purposes. The focus of most activity in academic environments is to use these new approaches to understand the fundamental basis of cellular and organismal biology. The primary objective of most industrial research, however, is to use such strategies to discover new drugs, or at least new lead compounds for drug discovery. These activities are not of course mutually exclusive, and indeed closer interactions between members of these two communities could bring substantial benefits to both parties.

The use of the vast libraries of new small molecules as 'chemical tools' to probe biological function and discover potential therapeutics is discussed in the reviews in this issue by Stockwell (page 846), and Lipinski and Hopkins (page 855). Using small molecules to probe biological systems is now often described as 'chemical genetics' or 'chemical genomics'²⁹. The enormous complexity of the biological milieu, again evident in Fig. 1, makes one of the ultimate goals of this approach — to discover a small molecule to modulate the function of every protein — an extremely challenging task, even in the light of the large arrays of chemical compounds that can be generated by combinatorial methods of ever-increasing sophistication. As well as the issues of diversity and specificity, cells may have evolved mechanisms to protect some of their most vital proteins from interference by small, extraneous molecules. Another major issue in chemical genetics concerns the quality of the data that are generated using various assay technologies; screening the same biological target with three different types of assay was recently found to give a set of hits that is consistent from assay to assay in only about 30% of cases³⁰. Although such a low level of consistency may not be very important for drug discovery, where the main objective is often simply to identify a number of active compounds, it can be debilitating if the objective is to chart the network of interactions within a biological organism. The quality of the chemical libraries and the reliability of screening techniques are still limiting factors in our knowledge of biological systems and their molecular diversity.

In addition to using the products of synthetic organic chemistry as tools to probe biological systems, new molecular tools based on other cellular components, such as DNA and RNA, are increasingly being developed. As Breaker discusses in a review in this issue (page 838), various RNA technologies are currently generating a great deal of interest. That RNA molecules play an important part in biological chemistry is well established, notably as the catalytic ribozymes that are involved in many important biological reactions, not least protein synthesis³¹. Moreover, RNA interference (RNAi), in which synthetic RNA fragments are designed to interfere with the normal expression of specific genes, is becoming an important tool for exploring gene function, as discussed at a recent Horizon Symposium, 'Understanding the RNAissance' (<http://www.horizonsymposia.com>), and reported

in ref. 32. In addition, aptamers — RNA molecules that form binding pockets for ligands with specificities and affinities similar to those of antibodies — are emerging as new probes of the functions of both large and small molecules. Aptamers that bind to particular targets can be engineered using *in vitro* evolution and amplification techniques. They can then be used as reagents to probe the roles of specific molecules in a given biological system. Furthermore, members of a previously neglected class of molecules, the oligosaccharides, are emerging as biological tools, now that efficient methods for sequencing and synthesizing these complex molecules are being developed³³. In addition to acting as probes of biological function and regulation, all these types of molecule are themselves becoming the focus of drug discovery efforts.

Future prospects

A rich array of data on the effects of small molecules on biological systems is accumulating, mainly from large-scale screening exercises (although the quality of this information is often less than optimal; see the review in this issue by Lipinski and Hopkins, page 855). Analysis of such databases, using the types of computational method pioneered by the flourishing bioinformatics community³⁴, should lead to major advances, both in our understanding of biological chemistry and in our ability to identify promising therapeutic compounds and therapeutic targets³⁵. Although progress is now being made in developing tools for mining chemical information, such progress is often limited by the difficulty in accessing much of the data of interest³⁶. Some estimates suggest that only about 1% of some types of chemical information are in the public domain. In contrast, the majority of many forms of biological data, from gene sequences to protein structures, is freely accessible to scientists in both academia and industry. One of the reasons for the inaccessibility of so much chemical information, in addition to the technical challenges of cataloguing and checking vast amounts of data, is concerned with issues of intellectual property. However, one can be optimistic that ways will be found to overcome the various hurdles to allow these resources to be used in the most effective ways possible.

With increasingly diverse, reliable and accessible databases of information about the effects of new chemical compounds on specific biochemical processes, we shall be able to understand much more about the nature of biologically relevant chemical space. In addition, we shall learn more about the types of compound that might make good drugs by analysing the behaviour of a much wider range of small molecules than the miserly number used by our bodies for so many purposes — from generating energy to building arsenals of macromolecules. In this regard, among the most exciting recent developments are efforts to generate public databases of chemical information³⁷, and the establishment by the US Government of Molecular Libraries Screening Centers. The latter initiative is designed to give public-sector researchers access to an initial library of around 500,000 small molecules for use in probing a diverse range of biological systems. These compounds may lead to new research tools and could aid the development of new drugs or the discovery of new applications for existing ones (see NIH Molecular Libraries Initiative, <http://nihroadmap.nih.gov>).

To exploit fully the emerging chemical tools and new methodologies in molecular and structural biology (for example, <http://www.nigms.nih.gov/psi/centers.html>), and so make the quantum leap in the efficiency of drug discovery that these developments promise, chemists must increasingly develop strong interactions with scientists from different disciplines. With such interdisciplinary collaborations it will be possible to embrace some of the grand challenges that exist in our quest to understand and manipulate the chemistry of life for the benefit of mankind. One of the greatest challenges must be to discover and understand what fraction of the universe of chemical space is used by living systems, and how much more could in principle be used to influence these systems. Progress in this area of science will lead to more efficient

strategies for drug discovery. And as such challenges are embraced, we shall very likely learn many of the secrets of how life began and evolved. □

doi:10.1038/nature03192

1. Fersht, A. R. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (W. H. Freeman, New York, 1999).
2. Bohacek, R. S., McMartin, C. & Guida, W. C. The art and practice of structure-based drug design: a molecular modelling perspective. *Med. Res. Rev.* **16**, 3–50 (1996).
3. Luria, S. E., Gould, S. J. & Singer, S. *A View of Life* (Benjamin/Cummings, Menlo Park, California, 1981).
4. Goto, S., Okuno, Y., Hattori, M., Nishioka, T. & Kanehisa, M. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.* **30**, 402–404 (2002).
5. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 806–921 (2001).
6. Ellis, R. J. & Minton, A. P. Join the crowd. *Nature* **425**, 27–28 (2003).
7. Hall, D. & Minton, A. P. Macromolecular crowding: qualitative and semiquantitative successes, quantitative challenges. *Biochim. Biophys. Acta* **1649**, 127–139 (2003).
8. Voet, D. & Voet, J. G. *Biochemistry* 2nd edn (Wiley, New York, 1995).
9. Goodsell, D. S. Inside a living cell. *Trends Biochem. Sci.* **16**, 203–206 (1991).
10. Westerhoff, H. V. & Palsson, B. O. The evolution of molecular biology into systems biology. *Nature Biotechnol.* **22**, 1249–1252 (2004).
11. Skär, J. & Coveney, P. V. Self-organization: the quest for the origin and evolution of structure. *Phil. Trans. R. Soc. Lond. A* **361**, 1047–1317 (2003).
12. Tsien, R. Y. Imagining imaging's future. *Nature Rev. Mol. Cell Biol.* **4**, SS16–SS21 (2003).
13. Medalia, O. *et al.* Macromolecular architecture in eukaryotic cells visualised by cryoelectron tomography. *Science* **298**, 1209–1213 (2002).
14. Kitano, H. Computational systems biology. *Nature* **420**, 206–210 (2001).
15. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **23**, 3–25 (1997).
16. Houghten, R. A. Parallel array and mixture-based synthetic combinatorial chemistry: tools for the next millennium. *Annu. Rev. Pharmacol. Toxicol.* **40**, 273–282 (2000).
17. Schreiber, S. L. Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science* **287**, 1964–1969 (2000).
18. Bleicher, K. H. *et al.* Hit and lead generation: beyond high-throughput screening. *Nature Rev. Drug Discov.* **2**, 369–378 (2003).
19. Feher, M. & Schmidt, J. M. Property distributions: differences between drugs, natural products and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **43**, 218–227 (2003).
20. Service, R. F. Surviving the blockbuster syndrome. *Science* **303**, 1796–1799 (2004).
21. Dickson, M. & Gagnon, J. P. Key factors in the rising cost of new drug discovery and development. *Nature Rev. Drug Discov.* **3**, 417–429 (2004).
22. Mullin, R. Drug Discovery. *Chem. Eng. News* **82**, 23–31 (2004).
23. Collins, F. S. *et al.* A vision for the future of genomics research. *Nature* **422**, 835–847 (2003).
24. Zambrowicz, B. P. & Sands, A. T. Knockouts model the 100 best-selling drugs — will they model the next 100? *Nature Rev. Drug Discov.* **2**, 38–51 (2003).
25. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nature Rev. Drug Discov.* **1**, 737–730 (2002).
26. Drews, J. Drug discovery: a historical perspective. *Science* **287**, 1960–1964 (2000).
27. Newman, D. J., Cragg, G. M. & Snader, K. M. Natural products as a source of new drugs over the period 1981–2002. *J. Nat. Prod.* **66**, 1002–1037 (2002).
28. Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740 (1997).
29. Schreiber, S. L. Chemical genetics resulting from a passion for synthetic organic chemistry. *Bioorg. Med. Chem.* **6**, 1127–1153 (1998).
30. Silis, M. A. *et al.* Comparison of assay technologies for a tyrosine kinase assay generates different results in high throughput screening. *J. Biomol. Screening* **7**, 191–214 (2002).
31. Steitz, T. A. & Moore, P. B. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. *Trends Biochem. Sci.* **28**, 411–418 (2003).
32. Novina, C. D. & Sharp, P. A. The RNAi revolution. *Nature* **430**, 161–164 (2004).
33. Seeberger, P. H. Automated carbohydrate synthesis to drive chemical glycomics. *Chem. Commun.* **10**, 1115–1121 (2003).
34. Buckingham, S. Bioinformatics: programmed for success. *Nature* **425**, 209–215 (2003).
35. Agrafiotis, D. K., Lobanov, V. S. & Salemme, F. R. Combinatorial informatics in the post-genomics era. *Nature Rev. Drug Discov.* **1**, 337–346 (2002).
36. Townsend, J. A. *et al.* Chemical documents: machine understanding and automated information extraction. *Org. Biomol. Chem.* **22**, 294–300 (2004).
37. Schreiber, S. L. The small-molecule approach to biology: chemical genetics and diversity-oriented organic synthesis make possible the systematic exploration of biology. *Chem. Eng. News* **81**, 51–61 (2003).
38. Richardson, J. S. & Richardson, D. C. Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl Acad. Sci. USA* **99**, 2754–2759 (2002).
39. Dobson, C. M., Gerrard, J. A. & Pratt, A. J. *Foundations of Chemical Biology* (Oxford Univ. Press, Oxford, 2001).
40. *Nature insight: protein misfolding* *Nature* **426**, 883–909 (2003).
41. Dobson, C. M. Protein folding and misfolding. *Nature* **426**, 884–890 (2003).
42. Williams, R. J. P. & Frausto da Silva, J. J. R. *The Natural Selection of the Chemical Elements* (Oxford Univ. Press, Oxford, 1997).
43. Westheimer, F. H. Why nature chose phosphates. *Science* **235**, 1173–1178 (1987).
44. Chin, J. W. *et al.* An expanded eukaryotic genetic code. *Science* **301**, 964–967 (2003).

Acknowledgements I thank the Wellcome and Leverhulme Trusts for their support through programme grants.

Competing interests statement The author declares that he has no competing financial interests.

Lessons from natural molecules

Jon Clardy & Christopher Walsh

Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA
(e-mail: jon_clardy@hms.harvard.edu)

Natural products have inspired chemists and physicians for millennia. Their rich structural diversity and complexity has prompted synthetic chemists to produce them in the laboratory, often with therapeutic applications in mind, and many drugs used today are natural products or natural-product derivatives. Recent years have seen considerable advances in our understanding of natural-product biosynthesis. Coupled with improvements in approaches for natural-product isolation, characterization and synthesis, these could be opening the door to a new era in the investigation of natural products in academia and industry.

In the past century, diverse classes of natural products have been isolated and their structures characterized. These discoveries, along with the elucidation of biological and biochemical mechanisms of therapeutic action, have been central to the work of organic and medicinal chemists. Natural products have been invaluable as tools for deciphering the logic of biosynthesis and as platforms for developing front-line drugs^{1,2}. For example, between 1981 and 2002, 5% of the 1,031 new chemical entities approved as drugs by the US Food and Drug Administration (FDA) were natural products, and another 23% were natural-product-derived molecules³. Natural products are still major sources of innovative therapeutic agents for infectious diseases (both bacterial and fungal), cancer, lipid disorders and immunomodulation⁴.

However, the complexity of many natural products can limit the scope for making chemical modifications to optimize their therapeutic use. Moreover, obtaining a renewable supply of active compounds from biological sources can be problematic. Nevertheless, as the recent multigram, total synthesis of the potent anti-cancer natural product discodermolide shows⁵, the increasing efficiency of synthetic organic chemistry has reduced the barrier posed by limited natural supply, even for materials with very complex structures.

Here, we examine some of the lessons from nature that remind us of the structural and mechanistic diversity of

natural small molecules, and evaluate the uncertain present and diminishing future interest for natural products as central players in the research strategies of pharmaceutical companies. We begin by describing the structural features of representative natural products of medicinal importance, their mechanisms of action and their biosynthesis, before turning to prospects for future discoveries.

Structural features of natural products

How do natural products compare with drugs?

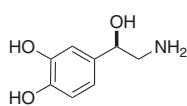
Figure 1a shows the structures of four natural products that have proved to be useful as drugs or leads: vancomycin^{6,7} (1), staurosporine⁸ (2), rapamycin⁹ (3) and Taxol¹⁰ (4). These have been used for the treatment of Gram-positive bacterial infections, as a lead indolecarbazole structure¹¹ for the inhibition of protein kinases at the ATP-binding site, for immunosuppression, and for cancer chemotherapy, respectively. For comparison, Fig. 1b shows the structures of four synthetic drug molecules that are in widespread use: Viagra¹² (5), Prozac¹³ (6), Lipitor¹⁴ (7), and Gleevec¹⁵ (8). These are used to treat erectile dysfunction, depression, hypercholesterolaemia and chronic myelogenous leukaemia, respectively. Each of the eight molecules has a well-defined biological target to which it binds with useful affinity, and all these targets are proteins, except for the peptidoglycan termini of bacterial cell walls (the target for

Box 1

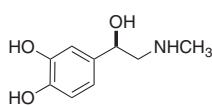
Signalling within and among organisms

Some important natural products with low molecular weights act with potency and specificity at protein receptors; for example, the low-molecular-weight amine neurotransmitters, derived from enzymatic decarboxylation of proteinogenic amino acids. These neurotransmitters have been outstanding platforms for natural-product-based drug design. Decarboxylation and subsequent oxidation of tyrosine generates the hormones and neurotransmitters noradrenaline (43) and adrenaline (44). Similar processing of tryptophan yields the neurotransmitter serotonin (45) and the hormone melatonin (46). Simple decarboxylation of histidine

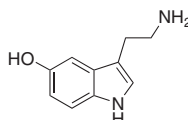
gives histamine (47), which has at least three well-characterized activities: (1) bronchoconstriction and vasodilation; (2) gastric-acid secretion; and (3) neurotransmission. These simple molecules have provided starting points for numerous small-molecule drugs. For example, seven out of ten anti-migraine medicines are based on serotonin³, several generations of α - and β -adrenergic drugs are generated from adrenaline scaffolds, and antihistamines (histamine receptor H1 and H2 selective antagonists) and selective serotonin re-uptake inhibitors (SSRIs) are some of the world's best-selling drugs.



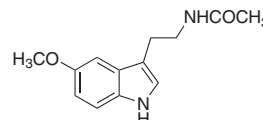
43 Noradrenaline



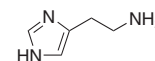
44 Adrenaline



45 Serotonin



46 Melatonin



47 Histamine

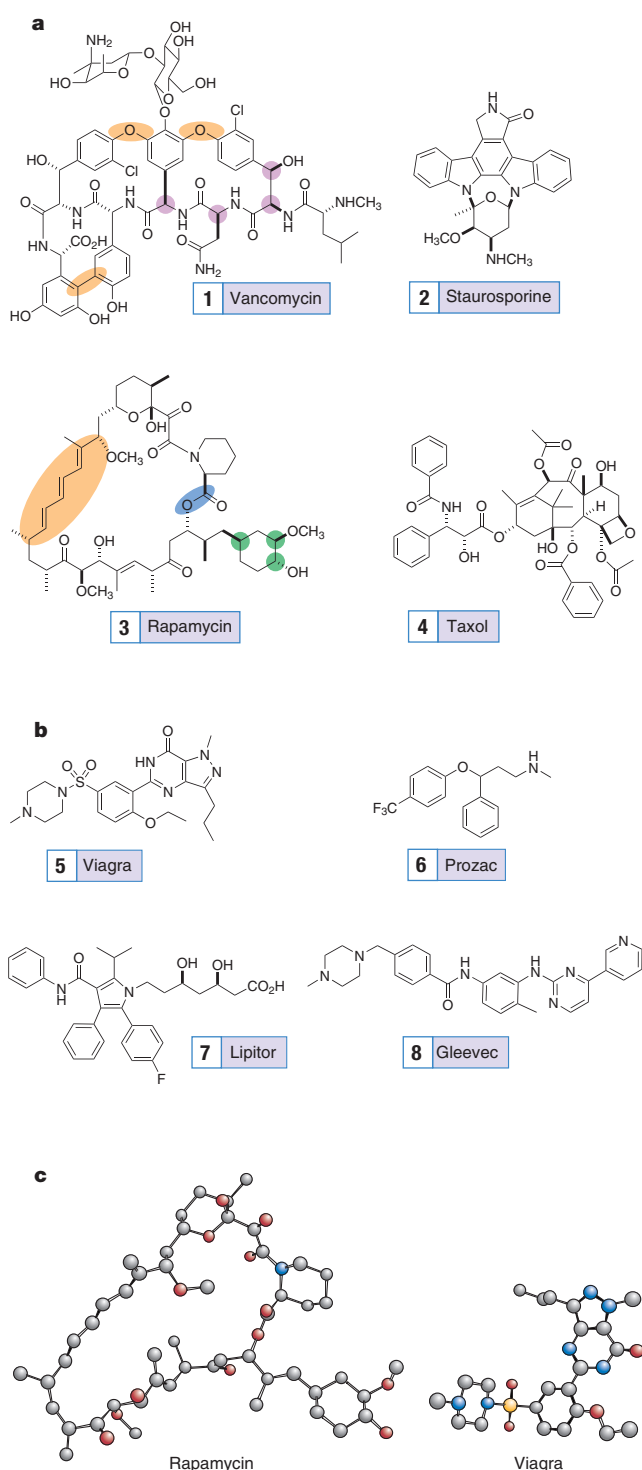


Figure 1 Medically significant natural products and synthetic molecules. **a**, Natural products. Vancomycin (**1**), an antibiotic for bacterial infections; staurosporine (**2**), a lead compound for the development of selective kinase inhibitors for cancer; rapamycin (**3**), a compound for immunosuppression; Taxol (**4**), an anti-cancer agent; **b**, Synthetic molecules. Viagra (**5**) for erectile dysfunction; Prozac (**6**) for depression; Lipitor (**7**) for hypercholesterolaemia; and Gleevec (**8**) for chronic myelogenous leukaemia. Natural products have strong conformational biases based on stereogenic centres (**1**, mauve circles), ether and ring fusions (**1**, yellow ovals), strategically placed substituents to select a single conformation (**3**, green circles), macrocyclization (**3**, blue oval), and conjugation (**3**, yellow oval). Staurosporine's (**2**) interlocking rings lead to a completely rigid core structure. **c**, Three-dimensional structural representations of rapamycin and Viagra.

vancomycin). Of the eight, only staurosporine is promiscuous in its recognition of protein targets; it binds to the ATP-recognition site of many protein kinases — a property that has limited its uses to a structural lead and a research tool¹¹.

Two-dimensional representations and three-dimensional images of these structures are shown in Fig. 1 to emphasize their architectural determinants. These comparisons highlight several general distinctions between natural-products and synthetic drugs/drug candidates. First, natural products typically have more stereogenic centres and more architectural complexity than synthetic molecules fashioned by medicinal chemists (Fig. 1), although several important natural products that act with potency and specificity at protein receptors have simple structures (Box 1). Second, natural products contain relatively more carbon, hydrogen and oxygen, and less nitrogen and other elements than synthetic medicinal agents. Third, many useful natural products have molecular masses in excess of 500 daltons and high polarities (greater water solubility), and therefore violate Lipinski's 'rule of five': this is a set of guidelines based on the characteristics of

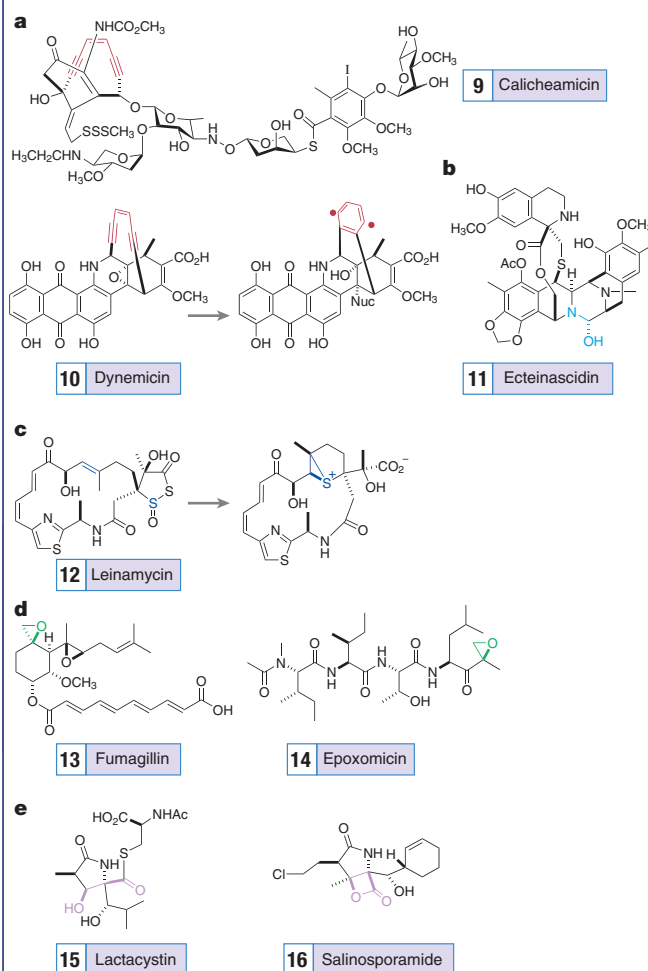


Figure 2 Natural products that exploit reactive functional groups. Compounds **9–16** all illustrate nature's ability to either mask or fine-tune the reactivity of functional groups. **a**, The enediyne group (red) in calicheamicin (**9**) and dynemicin (**10**) is activated to give a diradical intermediate that damages DNA (as shown for dynemicin). **b**, The carbolamine group in ecteinascidin (light blue; **11**) is converted to an iminium ion that reacts with DNA. **c**, The dithian-1,3-oxide group (dark blue) in leinamycin (**12**) is activated to form an episulphonium intermediate that alkylates DNA. **d**, Fumagillin (**13**) and epoxomicin (**14**) contain reactive epoxide groups (green) that trap proteases. **e**, The masked or explicit β -lactones (mauve) in lactacystin (**15**) and salinosporamide (**16**), respectively, target the proteasome.

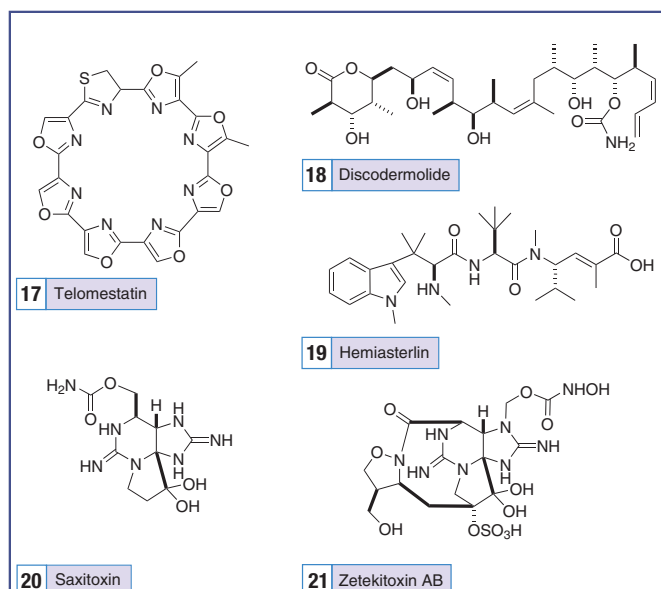


Figure 3 Natural products that exploit shape and polarity complementarity to biological targets. Biological targets include: G quartets for telomestatin (**17**); tubulin for discodermolide (**18**) and hemiasterlin (**19**); and ion channels for saxitoxin (**20**) and zetekitoxin AB (**21**).

known drugs that provide an indication of whether a given small molecule is likely to have the desired pharmacokinetic properties to be an oral drug (in terms of how it is absorbed, distributed, metabolized and eliminated by the body). All four synthetic drugs featured have a molecular mass lower than 500 daltons and can be orally administered.

Both the natural products and the synthetic drugs have strong conformational biases and constraints: examples include macrocyclizations (the formation of macrocyclic ring structures, such as that present in rapamycin shown in Fig. 1), fused-ring systems, ether crosslinks, extensive conjugation and strategically placed substituents that 'preorganize' them for populating conformers that bind to specific biological targets, in these cases enzymes and receptors (Fig. 1). The conformer restrictions and/or architectural rigidifications built into active molecules reflect the importance of minimizing the loss of entropy as molecules bind to biological targets. Avoiding such energy loss by preorganizing conformers to present complementary electrostatic, hydrogen-bonding and hydrophobic interactions with the protein targets allows these and other small molecules to retain sufficient binding energy to function as potent ligands. These are typically in the 10^{-7} to 10^{-9} M range of potency.

Lessons from natural-product functionalities

Natural products have been effective in teaching us about chemical functionality that is compatible with the aqueous milieu of biological microenvironments; the lessons learned have been both surprising and deep. Some notable examples of instructive natural products, all of which contain highly reactive functional groups or the precursors to such groups, are shown in Figs 2 and 3. The enediynes, including calicheamicin¹⁶ (**9**) and dynemicin¹⁷ (**10**) are among the most potent cytotoxic agents discovered, with 50%-effective dose ranges in cell-killing assays as low as 10^{-17} M — a nominal concentration in the range of one molecule per cell¹⁸. The unusual trisulphide in calicheamicin and the quinone in dynemicin are redox-activated triggers that initiate aromatization cascades leading to the formation of diradical intermediates that damage DNA (as shown for dynemicin in Fig. 2a). Nature frequently exploits such reactive functional groups in biologically active natural products. In ecteinascidin (**11**, Fig. 2b), a

carbinolamine is converted to an iminium ion that reacts with DNA to form a covalent adduct^{19,20}. In leinamycin (**12**, Fig. 2c), the dithian-1,3-oxide group in this anti-tumour agent is activated by a thiol to form an episulphonium intermediate that alkylates DNA¹⁷. Fumagillin (**13**) and epoxomicin (**14**) both use reactive epoxide groups to covalently trap proteases (Fig. 2d). Fumagillin's ability to selectively inhibit methionine aminopeptidase type 2 leads to the inhibition of angiogenesis (the formation of new blood vessels)²¹, and fumagillin-inspired compounds are being investigated as anti-cancer agents. Epoxomicin inhibits the degradation of proteins by the proteasome²², and related proteasome inhibitors are being developed for a variety of therapeutic uses. β -lactones, either masked as in lactacystin (**15**) or explicit as in salinosporamide (**16**, Fig. 2e) are also potent proteasome inhibitors^{23,24}. These examples illustrate nature's ability to either mask or finely tune the reactivity of labile functional groups so that a small molecule can retain the kinetic stability needed for it to reach and specifically inhibit biological targets by a covalent mechanism.

Of course, not all natural products work by covalent mechanisms; most employ the exquisite structural complementarity between a small molecule and its target. Telomestatin (**17**, Fig. 3), with its eight tandem heterocycles in a macrocyclic array, mimics the tetraguanine fragments (G quartets) found on telomeres²⁵. This mimicry allows telomestatin to be a nanomolar inhibitor of telomerase. Rapamycin (**3**, Fig. 1a) uses two different faces to bind two different proteins with nanomolar efficiency sufficient to disrupt a cytoplasmic signal transduction cascade²⁶. Discodermolide²⁷ (**18**) and hemiasterlin^{28,29} (**19**, Fig. 3) bind to tubulin, and both are exciting leads for cancer therapy. Natural products can also block ion channels, as illustrated by saxitoxin³⁰ (**20**) and zetekitoxin AB³¹ (**21**, Fig. 3).

The natural products in Fig. 1a are from traditional sources: soil microbes (vancomycin, **1**; staurosporine, **2**; and rapamycin, **3**) and plants (Taxol, **4**). Many of the natural products in Figs 2 and 3 are from nontraditional sources. Ecteinascidin (**11**) is from a small reef-dwelling tunicate found in the West Indies³². Discodermolide (**18**) is from a deep-water sponge and hemiasterlin (**19**) was found in two different sponges — one from South Africa, the other from Papua New Guinea. Saxitoxin (**20**) is produced by dinoflagellates (especially those producing 'red tides'), although it was traditionally isolated from filter-feeding shellfish that consumed the dinoflagellates. Its structural relative zetekitoxin AB (**21**) was isolated from the Panamanian golden frog, but its original producer is probably a microbe that is consumed by insects, which are in turn consumed by the frogs. This previously unexplored biological diversity coupled with modern analytical techniques and synthetic organic chemistry could lead to a new chapter of natural-products research, as is discussed in the section 'Discovery from new sources' below.

Understanding the functional-group arrays used by nature has informed synthetic- and medicinal-chemistry efforts about biomimetic strategies and isostere (shape-conserving) replacements. The synthetic molecules in Fig. 1b feature the design principles favoured by medicinal chemists: a high proportion of aromatic and heteroaromatic rings, few stereogenic centres, low molecular weights and a lack of chemical reactivity. In contrast, the enediyne anti-tumour antibiotic calicheamicin (**9**, Fig. 2a) is large (almost 1,400 daltons), devoid of core aromatic rings (until triggered by subsequent chemical reactions), loaded with stereogenic centres and highly reactive. Whether its potent biological properties can be exploited for anti-cancer therapy is not completely settled but an antibody-targeted-therapy approach Mylotarg, that takes advantage of its extraordinary cytotoxicity has been in the clinic since 2000 (see refs 33, 34).

Synthetic molecules are increasingly produced by combinatorial chemistry approaches, in which a common core is elaborated by attaching combinations of fragments to reactive sites on the core's periphery. An old, but still useful, template is the benzodiazepine core (**22**, Fig. 4a). In the construction of a synthetic combinatorial library based on the benzodiazepine skeleton (**22**), diversity elements (R_1 , R_2 and R_3) are attached to a common skeleton. If ten versions of

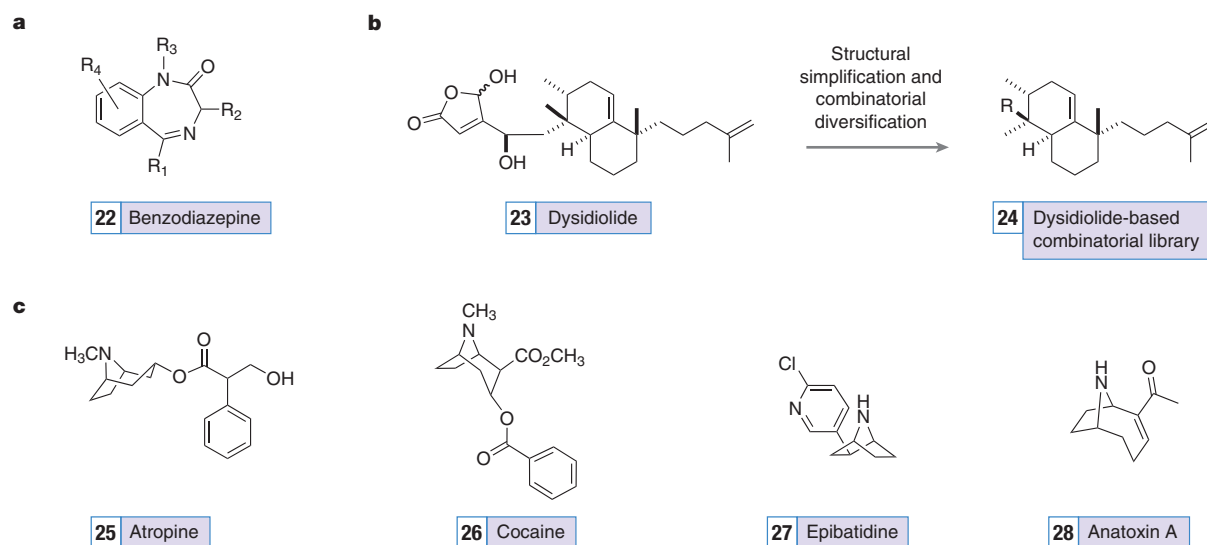


Figure 4 Template diversification. **a**, The benzodiazepine core is a common template for synthetic diversification because the groups indicated here as R_1 , R_2 , R_3 and R_4 can be varied widely. **b**, Dysidiolide (**23**) has been used as a template for a natural-products-based diversity library. The native structure was simplified and a single diversity element was used to create the library (**24**).

c, Compounds **25–28** illustrate natural template diversification. Atropine (**25**) and cocaine (**26**) are plant alkaloids with mydriatic and local anaesthetic properties, respectively. Epibatidine (**27**) is a non-opioid analgesic isolated from the skin of an Ecuadoran poison frog, and anatoxin A (**28**) is the Very Fast Death Factor produced by cyanobacteria.

each diversity element are used, the library contains 1,000 different molecules, each with a different combination of R_1 , R_2 and R_3 . Nature uses similar strategies, especially the oxidative elaboration of a central core followed by capping reactions (discussed in the section ‘Re-engineering of biosynthetic pathways’ below). Several natural-product-like combinatorial libraries have been synthesized³⁵; a library based on dysidiolide (**23**, Fig. 4b), a potent phosphatase inhibitor from a marine sponge³⁶, and summarized in structure **24**, is typical³⁷. In this case, the library construction involved the diversification of a single element on a single scaffold (**24**); even with these apparent limitations, the library contained potent phosphatase inhibitors³⁷. Successes with combinatorial libraries based on natural-product templates argue that natural products, which have been honed by their evolutionary history for biological activity, are excellent starting points for structural diversification³⁷. Combinatorial biosynthesis (which is discussed in the section ‘Re-engineering of biosynthetic pathways’ below) uses the manipulation of biosynthetic machinery to accomplish much the same goal, but with greater control over core elements.

Nature also dramatically varies the core size and stereochemistry of molecules, as the series atropine (**25**), cocaine (**26**), epibatidine (**27**) and anatoxin A (**28**) illustrates (Fig. 4c). Diversity-oriented synthesis^{38,39}, which combines the strengths of combinatorial multiplexing and core variability, is emerging as a powerful technique for finding biologically active small molecules⁴⁰.

Advantages and constraints of nature’s biosynthetic strategy

Natural products can be divided into several structural classes: polyketides, nonribosomal peptides (NRPs), terpenes, alkaloids and many others. Products are classed according to shared scaffolding elements, which in turn reflect the strategies for their assembly by pathways of biosynthetic enzymes in the producer organisms.

Most classical small molecules from nature are secondary metabolites — products from conditional pathways that are turned on in a particular context or situation. These include metabolites made during starvation (for example, carbapenem antibiotics produced by *Pseudomonas* bacteria), in development (for example, antibiotics made when *Streptomyces* enter cellular differentiation

pathways), and signalling (such as quorum-sensing molecules biosynthesized at particular culture densities of microbes)⁴¹.

The building blocks for natural products are most often the monomer constituents (amino acids for nonribosomal peptides; acyl-CoA thioesters for polyketides; isoprenyl-pyrophosphates for terpenes) of primary metabolic pathways, which are shunted into the secondary pathways when a particular metabolic channel is opened. When monomers dedicated to secondary metabolic pathways are required, such as 4-OH-phenylglycine and 3,5-(OH)₂-phenylglycine for vancomycin (Fig. 1a) and methoxymalonyl CoA for some polyketide initiations, they are produced by a ‘just-in-time’ cellular-inventory strategy⁴². To this end, biosynthetic gene clusters for non-ribosomal peptides or NRPs (such as vancomycin) and polyketides (such as rapamycin in Fig. 5) contain both genes for the assembly-line enzymes and genes for enzymes to make the dedicated monomers needed for the assembly lines to run⁴³. A third set of clustered genes typically encodes enzymes that tailor the nascent products released from assembly lines, most notably for glycosylation and oxidation: these two modifications are often required to make the product biologically active⁴⁴. The gene clustering allows coordinated regulation and inventory control of both enzyme catalysts and small-molecule building blocks. The enzyme catalysts are needed to run the secondary pathways comprising 20 to 40 steps that turn out the finished natural products.

The simple monomers are used in sets of iterative condensations; linear intermediates are built up by a single type of chemistry. For example, for terpene and isoprenoid natural products, the fundamental chain-elongation step is C-alkylation enzyme catalysis, which adds a C₅-isoprene unit to the end of a growing chain by means of allylic carbonium ion chemistry. The growing chain is held in the microenvironment of the oligomerizing enzymes that control foldamer conformation. This in turn dictates cyclization patterns, such as in Taxol or polycyclic triterpene assembly. In NRP- and polyketide-chain buildup, both the growing chain and the incoming monomer are tethered covalently to the enzyme as thioesters. For NRPs, the iterative elongation step is amide-bond formation, whereas in polyketide-chain growth it is Claisen-type C–C bond formation to the β-keto-acyl thioester products.

How is so much structural diversity generated in these three classes of natural products, which are produced from a limited pool of simple primary metabolites? The general answers are incomplete processing and/or active tailoring of the initial intermediates during chain elongation, acyclic foldamer control for regiospecific cyclization reactions and post-elongation tailoring and maturation by enzyme action. In terpenes, foldamer control and the placement of basic side chains in the terpene cyclase active sites controls the location and size of cation-mediated cyclizations^{45,46}. In NRP-assembly lines, cysteinyl, seryl and threonyl side chains can be regiospecifically cyclized, dehydrated and oxidized to create thiazoles and oxazoles during elongation. In multimodular polyketide assembly lines, the initial β -keto-acyl thioesters from Claisen condensation can be processed all the way to β -CH₂ methylene groups or can accumulate as β -keto, β -hydroxy or $\alpha\beta$ -olefinic intermediates^{47,48}. Where full-length peptidyl thioesters or full-length polyketidyl thioesters have been assembled on the most downstream way stations of NRP- and polyketide-assembly lines, chain release can occur through hydrolysis. Alternatively, chain release can occur through an intramolecular regiospecific cyclization from a nucleophilic -OH or -NH in the chain to form a macrolactone or macrolactam. Intramolecular release results in a macrocycle that builds in conformational constraints⁴⁹.

The biosynthesis of the immunosuppressive drug rapamycin (3) illustrates how structural diversity is generated from simple building

blocks⁵⁰. As noted in Fig. 5, this is predominantly a polyketide natural product with a dihydroxycyclohexenyl CoA as a starting building block, and seven equivalents of malonyl CoA and seven equivalents of methylmalonyl CoA as the elongating monomers. One amino acid is incorporated; in this case L-pipecolate, which is derived from a dedicated enzymatic cyclization of the primary metabolite lysine. The order in which these four classes of monomer are incorporated is determined by the order of the 15 modules in the enzymatic assembly line. Figure 5 shows how the single nonribosomal peptide synthase (NRPS) module is at the end, suggesting that pipecolate is the last monomer to be incorporated. The linear acyl-S-enzyme intermediate that is proposed to undergo capture by intramolecular cyclization to yield the 30-member macrolactam is also shown. The nascent macrocyclic product is then tailored by a series of enzymatic methylations and oxidation/oxygenation steps to yield rapamycin.

An analogous but distinct logic is used in the assembly of the enediynes cores by polyketide synthase assembly lines, which are then followed by tailoring reactions^{51,52}. Altogether, 55 enzymatic reactions are used to combine five classes of building block (acetyl CoA, malonyl CoA, tyrosine, chorismate and glucose) to give the enediyne C-1027 (ref. 53).

Tailoring reactions to control oxidation states

All the linear chain-elongation steps in polyketide and NRP monomer assembly occur as thioesters, and without any protecting

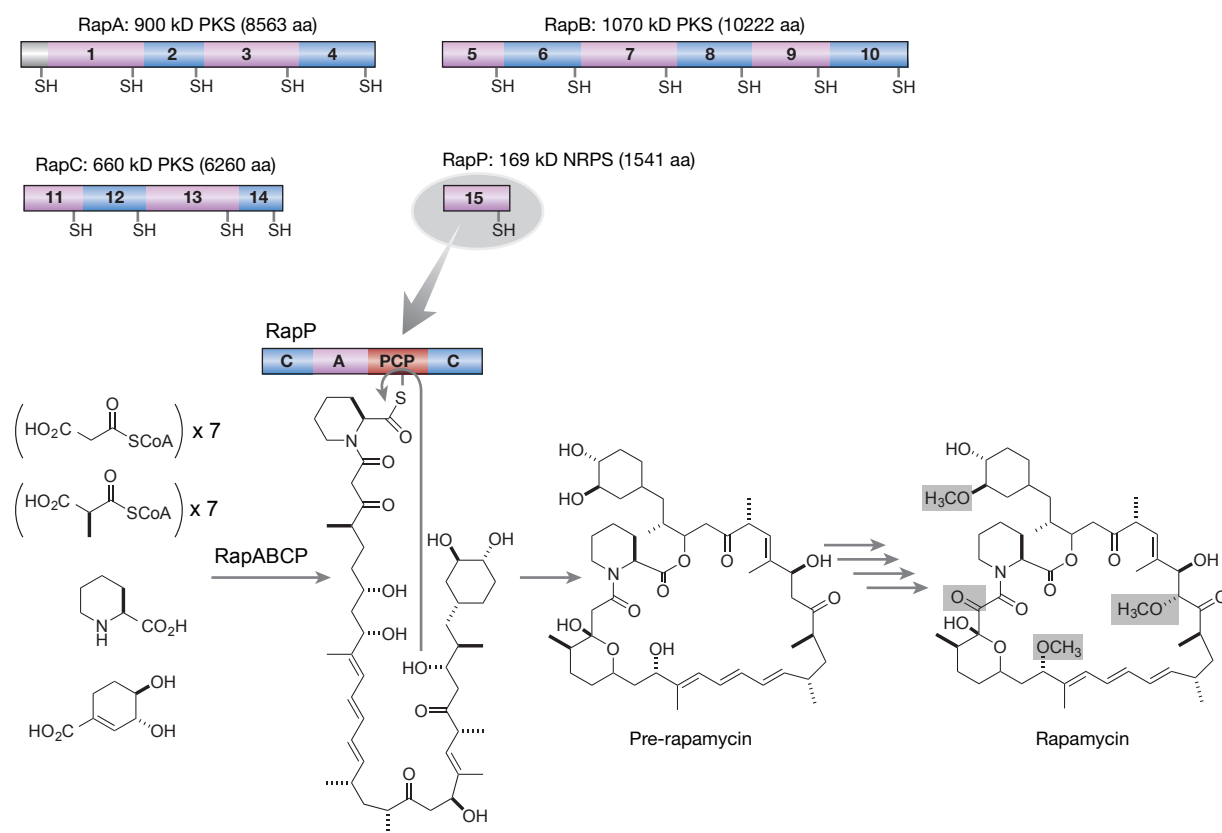
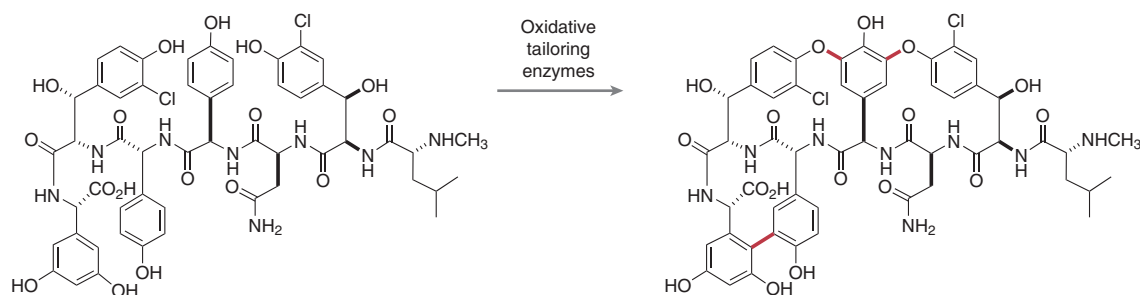


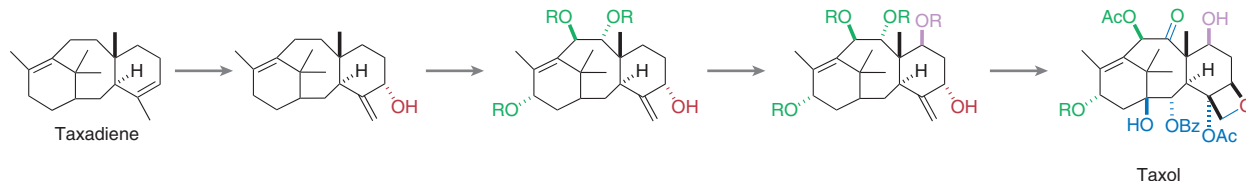
Figure 5 Biosynthesis of natural products. The rapamycin synthase assembly line consists of four multimodular proteins (RapA, RapB, RapC and RapP). Fourteen polyketide synthase modules are distributed in RapA–C and the fifteenth, a nonribosomal peptide synthase module (NRPS), comprises the RapP protein. RapA–C comprise the three-subunit assembly-line machinery for the polyketide-chain initiation and elongation. Each of the 15 modules has a carrier-protein domain (peptidyl carrier protein, PCP in RapP). This is post-translationally modified with a phosphopantetheinyl arm containing a terminal cysteine on which the elongating acyl chains are assembled. The most downstream acyl

intermediate is shown on the PCP domain of RapP as it undergoes an intramolecular cyclization, thought to be catalysed by the second condensation domain (C) of RapP. The first C domain makes the acyl–N linkage to the pipecolyl moiety of the acyl chain, while the adenylation domain (A) selects, activates and incorporates the pipecolyl moiety. All the atoms of pre-rapamycin come from the four building blocks malonyl CoA, methylmalonyl CoA, pipecolate and dihydroxycyclohexenyl CoA, as shown. After cyclo-release from the assembly line, pre-rapamycin undergoes a series of oxidative and O-methylation-tailoring steps to yield the final product: rapamycin.

a



b



c

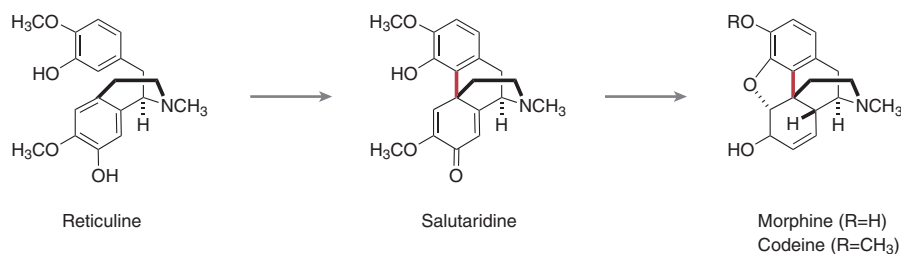


Figure 6 The role of oxidation in the construction of natural products. **a**, The oxidative tailoring of vancomycin by three haem-containing proteins introduces aryl ether (C–O) bonds and aromatic (C–C) crosslinks (shown in red) that rigidify the vancomycin skeleton. **b**, The spectacular series of oxidations that convert taxadiene to Taxol. Eight oxygen atoms are introduced into the scaffold by cytochrome P450 mono-

oxygenases, and these are further modified into carbonyl, ether or ester links. The intermediates shown have been identified, but not all the responsible enzymes have been characterized; some of the transformations require more than one enzyme^{57,62}. **c**, A key step in the biosynthesis of morphine and other opium alkaloids involves the oxidative coupling of two phenol radicals to form the key bond shown in red.

groups to mask the reactivity of ketones, alcohols and olefins. In particular, the alcohol side chains in the nascent products from a polyketide assembly line come from the carbonyl groups of malonyl or methylmalonyl monomers. However, additional hydroxyl groups (derivatives of oxygen) are introduced by tailoring mono-oxygenases that act with regio- and stereospecificity. A spectacular case of post-assembly-line oxidative tailoring logic occurs in vancomycin biosynthesis, where three haem proteins, all encoded in the biosynthetic gene cluster, act in temporal and regiochemical sequence to introduce the 4–6 and 2–4 aryl ether⁵⁴, and 5–7 (C–C) crosslinks⁵⁵ in the aglycone scaffold (Fig. 6a). These crosslinks generate the rigid architecture necessary for high-affinity recognition of the *N*-acyl-D-Ala-D-Ala termini of bacterial peptidoglycan strands.

Baldwin noted at a recent Horizon Symposium (<http://www.horizonsymposia.com>) that the oxygenative maturation of the taxane skeleton to Taxol reveals a comparable enzymatic strategy of assembling the taxane scaffold in a reduced oxidation state and then conducting regiospecific and stereospecific enzymatic oxidations (Fig. 6b). The initial cyclization product from the C₂₀ isoprenoid geranylgeranyl pyrophosphate is taxa-4(5),11(12)-diene (ref. 56). This intermediate undergoes eight specific hydroxylations by cytochrome P450 mono-oxygenases^{57,58}. Four of the newly introduced hydroxyls are then enzymatically acylated, allowing precisely controlled oxidation on the periphery of the tetracyclic scaffold.

A third example of late-stage redox tailoring is found in the reticuline to salutaridine to morphine pathway (Fig. 6c). These examples of late-stage redox tailoring contrast with a chemist's approach towards total synthesis. Here, fragments are prepared using a convergent, not a linear, approach; the fragments have protecting groups that can be orthogonally manipulated, and the desired final oxidation states are mainly built into the strategy of fragment construction. As a result, synthetic chemists have a much larger set of building blocks with which to carry out their convergent strategies. Despite this, the regio- and stereospecific hydroxylation of related carbon centres in complex molecular scaffolds remain synthetic challenges for which the chemist cannot readily mimic the natural enzymatic process.

Re-engineering of biosynthetic pathways

The burgeoning database of microbial genomes has led to the cataloguing of hundreds of gene clusters that encode polyketides, NRPs and hybrid polyketide–NRP natural products⁵⁹. The coding logic can be deciphered in some cases to make good predictions of what dedicated metabolites will be used as monomers for the assembly lines, what the structures of advanced intermediates will be, and whether post-assembly-line tailoring steps, such as methylations, acylations, glycosylations and oxidations (including hydroxylations), are encoded. These create a set of catalytic-part lists for engineering new polyketide, peptide and hybrid ‘unnatural’ natural products by

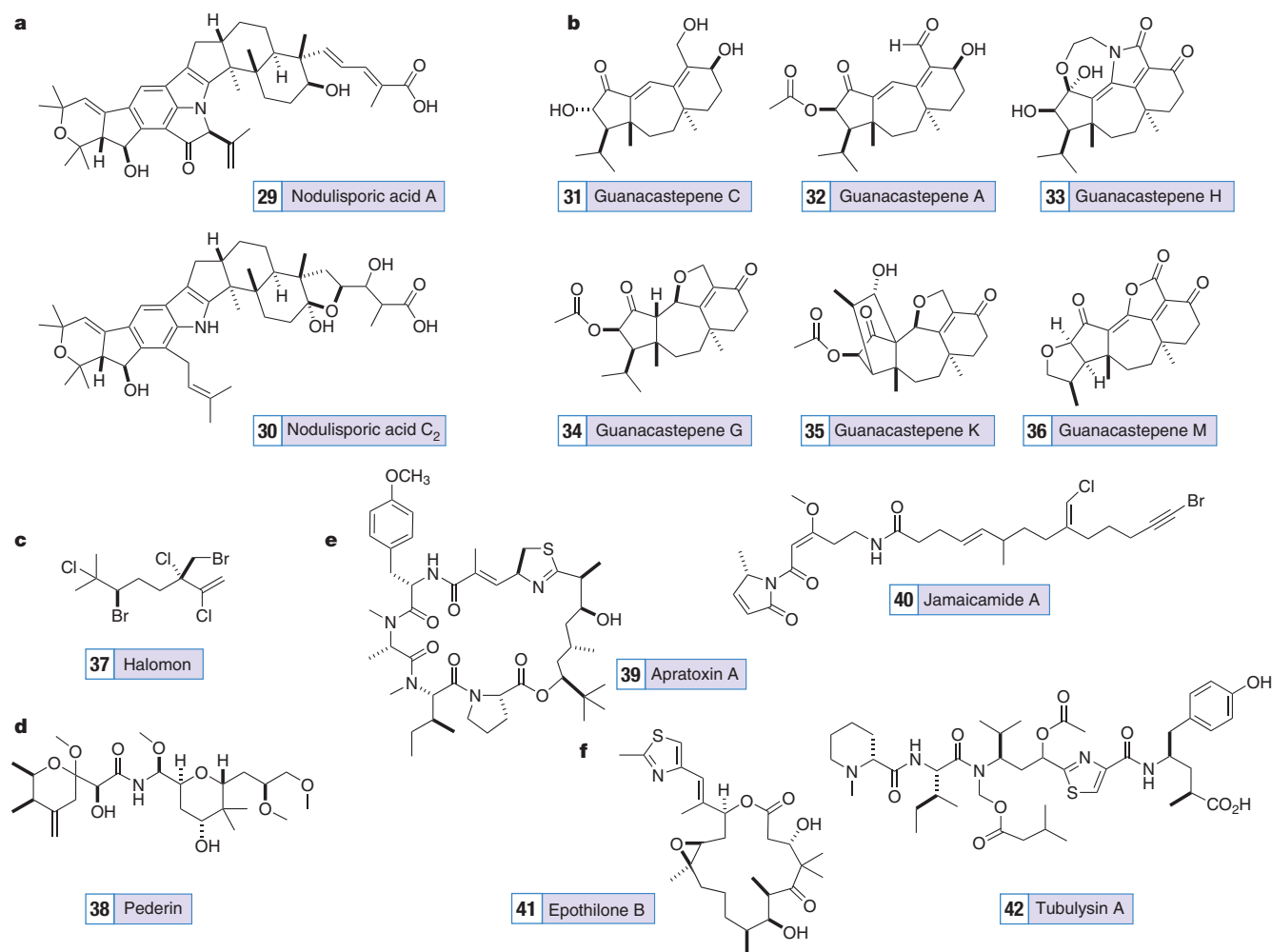


Figure 7 Recent natural products obtained from nontraditional sources. **a**, The nodulisporic acids (**29**, **30**) and **b**, the guanacastepenes (**31–36**) are from endophytic fungi, the large group of fungi that live inside higher plants; **c**, halomon (**37**) is from a

red alga; **d**, pederin (**38**), which was long believed to be an insect metabolite, is produced by bacteria; **e**, apratoxin (**39**) and jamaicamide (**40**) are from marine cyanobacteria; **f**, epothilone (**41**) and tubulysin (**42**) are from myxobacteria.

domain deletions, mutations and swaps^{43,60}. Glycosylation variants in anthracycline anti-tumour molecules⁶¹ and in glycopeptide antibiotics of both the vancomycin and teicoplanin scaffolds have been reported *in vitro*⁶², whereas engineered glycosylations of tetracyclic aromatic polyketides have been conducted *in vivo*⁶³. Dozens of variants of the deoxyerythronolide scaffold in the erythromycin family have been produced by re-engineering up to three catalytic domains at a time in the three-subunit deoxyerythronolide B (DEB) synthase⁶⁴. This is followed by tailoring glycosylation⁶⁵.

With current gene-synthesis technology, it is possible to make assembly lines, for example, for DEB synthase, using dozens of designed restriction sites. These allow chemists to swap a given domain or module with synthetic or natural genetic variants from any other assembly line. Gene-shuffling methodologies similarly increase diversity, so large libraries of variant synthases for polyketides, NRPs and polyketide–NRP hybrids (rapamycin, FK506, bleomycin and epothilones, for example) can probably be constructed and their yields of new products determined by structure-based and/or activity-driven screens. The feeding of alternate monomers into native and engineered assembly lines also leads to new natural-product variants⁶⁶.

The engineering of *Escherichia coli* to express the key taxadiene intermediate in the assembly of Taxol⁶⁷ and the sesquiterpene amorphadiene — a precursor to the anti-malarial agent

artemisinin⁶⁸ — have been described, thus indicating that the reconstruction of regiospecific terpene cyclization machinery can also be accomplished.

Discovery from new sources

Each time chemists are able to access new swathes of biological diversity, new — often strikingly new — natural products are discovered. Indeed, what are currently the most interesting natural products come mainly from recently accessed biota. The realization that there was a large, and largely unexplored, group of fungi (endophytic fungi) living inside higher plants led to focused discovery efforts in both industrial and academic laboratories. The nodulisporic acids (**29** and **30**, Fig. 7a) were discovered in an endophytic fungus from Hawaii⁶⁹. The guanacastepenes (**31–36**, Fig. 7b) were isolated from an endophytic fungus from Costa Rica using an antibiotic assay⁷⁰. The guanacastepenes provide an elegant illustration of nature's ability to use late-stage redox reactions to re-model core structures and produce a suite of diverse molecular skeletons. This core diversity differs from a typical synthetic combinatorial library featuring only peripheral modifications.

Exploration of the marine environment has also had a profound effect on natural-products chemistry. Early investigations focused on highly halogenated metabolites such as halomon⁷¹ (**37**, Fig. 7c) from a red alga, but many of the most structurally intriguing and biologically

potent molecules, such as discodermolide (**18**) and hemiasterlin (**19**), have come from sponges. Sponges are full of bacterial symbionts, and many sponge metabolites probably have bacterial origins. An interesting preliminary study has shown that pederin (**38**, Fig. 7d) — a well-known insect metabolite with a very similar structure to that of several sponge metabolites — has a bacterial origin⁷².

Other productive new sources include cyanobacteria, as represented by apratoxin A⁷³ (**39**), a potent cytotoxin with an unknown mechanism of action, and jamaicamide⁷⁴ (**40**), a potent neurotoxic sodium-channel blocker (Fig. 7e). Both apratoxin A and jamaicamide were isolated from the cyanobacteria *Lyngbya majuscula* (one strain from Guam, the other from Jamaica). Myxobacteria (gliding bacteria) have also been excellent producers of structurally interesting and biologically active natural products. Derivatives of epothilone⁷⁵ (**41**) from *Sorangium cellulosum* are currently being used in cancer trials, and tubulysins⁷⁶ (**42**) from *Archangium gephyra* are potent tubulin disruptors with potential anti-cancer activity (Fig. 7f).

As the examples from cyanobacteria and myxobacteria suggest, there are still many natural products to be discovered from bacteria. This biosynthetic potential is not surprising because bacteria encompass the main pool of genetic diversity on the planet; they interact with their surroundings, competitors and community members through small molecules, and they are largely unexplored. Fewer than 1% of the bacteria on Earth, and probably fewer than 0.1%, have ever been cultured⁷⁷. Most bacteria live in microbial communities where the members are mutually dependent on each other, and because current culturing practices select for strains that can live on their own, most bacteria are not cultured⁷⁸. Several approaches to dealing with uncultured bacteria have been proposed, including reconstituting the natural communities⁷⁸ and capturing biosynthetic gene clusters directly from DNA taken from the environment^{79,80}. Finally, the wealth of bacterial genomic data now available emphasizes that there are many biosynthetic gene clusters in culturable bacteria for which no associated small molecule can be isolated. Natural-product production is a highly regulated process and these cryptic pathways are not turned on under standard culturing conditions. A genomics-guided approach to discovering, sequencing and expressing these pathways has been described⁸¹.

Conclusions

The inventory of natural molecules remains incomplete, and discoveries of new structures and functions are likely to continue as underexplored sources of natural products are more systematically evaluated. The functional-group diversity and architectural platforms engineered into natural products during biosynthesis continue to provide lessons for synthetic and medicinal chemists in their strategies for making biologically active mimics, and provide selective ligands for cellular targets. Deciphering the molecular logic of biosynthetic enzymes and pathways, as monomers are assembled and nascent products tailored, has opened up practical approaches to re-engineering assembly lines to create unnatural variants of natural products. The molecular scaffolds created and used in nature are likely to persist as central design elements in subsequent generations of synthetic and semi-synthetic ligands that could become therapeutic agents for receptors, enzymes and ion channels.

Finally, although there has been a trend within the pharmaceutical industry to downscale efforts in natural-products research in recent years, careful reconsideration of this area could change this. Several problems with natural products that influenced the original company decisions to withdraw from the field (such as the challenges associated with identifying the active components from natural-product extracts that typically contain several compounds) are being addressed by technological advances. For example, the throughput of methods for compound purification and identification has increased. It seems clear that there is still great potential for accessing therapeutically relevant chemical diversity from nature — in particular, from the many organisms that have not yet been cultured. A revival in

interest in using natural products in early-stage drug discovery could be exactly what is needed to boost pharmaceutical output. □

doi:10.1038/nature03194

- Newman, D. J., Cragg, G. M. & Snader, K. M. The influence of natural products upon drug discovery. *Nat. Prod. Rep.* **17**, 215–234 (2000).
- Newman, D. J., Cragg, G. M., Holbeck, S. & Sausville, E. A. Natural products and derivatives as leads to cell cycle pathway targets in cancer chemotherapy. *Curr. Cancer Drug Targets* **2**, 279–308 (2002).
- Newman, D. J., Cragg, G. M. & Snader, K. M. Natural products as sources of new drugs over the period 1981–2002. *J. Nat. Prod.* **66**, 1022–1037 (2003).
- Altmann, K.-H. Microtubule-stabilizing agents: a growing class of important anticancer drugs. *Curr. Opin. Chem. Biol.* **5**, 424–431 (2001).
- Mickel, S. J. *et al.* Large-scale synthesis of the anti-cancer marine natural product (+) – discodermolide. Part 1: Synthetic strategy and preparation of a common precursor. *Org. Proc. Res. Dev.* **8**, 92–100 (2004).
- Sheldrick, G. M., Jones, P. G., Kennard, O., Williams, D. H. & Smith, G. A. Structure of vancomycin and its complex with acetyl-D-alanyl-D-alanine. *Nature* **271**, 223–225 (1978).
- Williamson, M. P. & Williams, D. H. Structure revision of the antibiotic vancomycin. Use of nuclear Overhauser effect difference spectroscopy. *J. Am. Chem. Soc.* **103**, 6580–6585 (1981).
- Furasaki, A. *et al.* The crystal and molecular structure of staurosporine, a new alkaloid from a *Streptomyces* strain. *J. Chem. Soc. Chem. Commun.* 800–801 (1978).
- Sehgal, S. N., Baker, H. & Vezina, C. Rapamycin (AY-22,989), a new antifungal antibiotic. II. Fermentation, isolation and characterization. *J. Antibiot.* **28**, 727–732 (1975).
- Wani, M. C., Taylor, H. L., Wall, M. E., Coggon, P. & McPhail, A. T. The isolation and structure of taxol, a novel antileukemic and antitumor agent. *J. Am. Chem. Soc.* **93**, 2325–2327 (1971).
- Pindur, U., Kim, Y. S. & Mehrabani, F. Advances in indolo[2,3-a]carbazole chemistry: design and synthesis of protein kinase C and topoisomerase I inhibitors. *Curr. Med. Chem.* **6**, 29–69 (1999).
- Carson, C. C. 3rd Sildenafil: a 4-year update in the treatment of 20 million erectile dysfunction patients. *Curr. Urol. Rep.* **4**, 488–496 (2003).
- Wong, D. T. & Bymaster, F. P. Development of antidepressant drugs. Fluoxetine (Prozac) and other selective serotonin uptake inhibitors. *Adv. Exp. Med. Biol.* **363**, 77–95 (1995).
- Roth, B. D. The discovery and development of atorvastatin, a potent novel hypolipidemic agent. *Prog. Med. Chem.* **40**, 1–22 (2002).
- Kurzrock, R., Kantarjian, H. M., Druker, B. J. & Talpaz, M. Philadelphia chromosome-positive leukemias: from basic mechanisms to molecular therapeutics. *Ann. Intern. Med.* **138**, 819–830 (2003).
- Lee, M. D., Dunne, T. M., Chang, C. C., Morton, G. O. & Borders, D. B. Calicheamicins, a novel family of antitumor antibiotics. *J. Am. Chem. Soc.* **109**, 3464–3466 (1987).
- Konishi, M. *et al.* Crystal and molecular structure of dynemicin A: a novel 1,5-di-3-ene antitumor antibiotic. *J. Am. Chem. Soc.* **112**, 3715–3716 (1990).
- Thorson, J. S. *et al.* Understanding and exploiting nature's chemical arsenal: the past, present and future of calicheamicin research. *Curr. Pharm. Des.* **6**, 1841–1879 (2000).
- Scotto, K. W. ET-743: more than an innovative mechanism of action. *Anticancer Drugs* **13** (Suppl. 1), S3–6 (2002).
- Aune, G. J., Furuta, T. & Pommier, Y. Ecteinascidin 743: a novel anticancer drug with a unique mechanism of action. *Anticancer Drugs* **13**, 545–555 (2002).
- Liu, S., Widom, J., Kemp, C. W., Crews, C. M. & Clardy, J. Structure of human methionine aminopeptidase-2 complexed with fumagillin. *Science* **282**, 1324–1327 (1998).
- Meng, L., Kwok, B. H., Sin, N. & Crews, C. M. Eponemycin exerts its antitumor effect through the inhibition of proteasome function. *Cancer Res.* **59**, 2798–2801 (1999).
- Fenteany, G. *et al.* Inhibition of proteasome activities and subunit-specific amino-terminal threonine modification by lactacystin. *Science* **268**, 726–731 (1995).
- Feling, R. H. *et al.* Salinosporamide A: a highly cytotoxic proteasome inhibitor from a novel microbial source, a marine bacterium of the new genus *salinispora*. *Angew. Chem. Int. Edn Engl.* **42**, 355–357 (2003).
- Kim, M. Y., Gleason-Guzman, M., Izbicka, E., Nishioka, D. & Hurley, L. H. The differential biological effects of telomestatin and TMAPPyP4 can be attributed to their selectivity for interaction with intramolecular and intermolecular G-quadruplex structures. *Cancer Res.* **63**, 3247–3256 (2003).
- Choi, J., Chen, J., Schreiber, S. L. & Clardy, J. Structure of the FKBP12-rapamycin complex interacting with the binding domain of human FRAP. *Science* **273**, 239–242 (1996).
- Gunasekera, S. P., Gunasekera, M., Longley, R. E. & Schulte, G. K. Discodermolide: a new bioactive polyhydroxylated lactone from the marine sponge *Discodermia dissoluta*. *J. Org. Chem.* **55**, 4912–4915 (1991).
- Talpir, R., Benayahu, Y., Kashman, Y., Pannell, L. & Schleyer, M. Hemiasterlin and geodiamolide TA: two new cytotoxic peptides from the marine sponge *Hemiasterella minor*. *Tetrahedron Lett.* **35**, 4453–4456 (1994).
- Anderson, R. J., Coleman, J. E., Andersen, R. J. & Roberge, M. Cytotoxic peptides hemiasterlin, hemiasterlin A and hemiasterlin B induce mitotic arrest and abnormal spindle formation. *Cancer Chemother. Pharmacol.* **39**, 223–226 (1997).
- Schantz, E. J. *et al.* The structure of saxitoxin. *J. Am. Chem. Soc.* **93**, 7344–7345 (1975).
- Yotsu-Yamashita, M. *et al.* The structure of zetekitoxin AB, a saxitoxin analog from the Panamanian golden frog *Atelopus zetekii*: a potent sodium-channel blocker. *Proc. Natl. Acad. Sci. USA* **101**, 4346–4351 (2004).
- Rinehart, K. L. *et al.* Ecteinascidins 729, 743, 745, 759A, 759B and 770: potent antitumor agents from the Caribbean tunicate *Ecteinascidia turbinata*. *J. Org. Chem.* **55**, 4512–4515 (1990).
- Boghaert, E. R. *et al.* Antibody-targeted chemotherapy with the calicheamicin conjugate hu3S193-N-acetyl gamma calicheamicin dimethyl hydrazide targets Lewisy and eliminates Lewisy-positive human carcinoma cells and xenografts. *Clin. Cancer Res.* **10**, 4538–4549 (2004).
- Dijoseph, J. F. *et al.* Antibody-targeted chemotherapy with CMC-544: a CD22-targeted immunoconjugate of calicheamicin for the treatment of B-lymphoid malignancies. *Blood* **103**, 1807–1814 (2004).
- Boldi, A. M. Libraries from natural product-like scaffolds. *Curr. Opin. Chem. Biol.* **8**, 281–286 (2004).
- Gunasekera, S. P., McCarthy, P. J., Kelly-Borges, M., Lobkovsky, E. & Clardy, J. Dysidiolide: a novel protein phosphatase inhibitor from the Caribbean sponge *Dysidea etheria* de Laubenfels. *J. Am. Chem. Soc.* **118**, 8759–8760 (1996).

37. Brohm, D. *et al.* Natural products are biologically validated starting points in structural space for compound library development: solid-phase synthesis of disidiolide-derived phosphatase inhibitors. *Angew. Chem. Int. Edn Engl.* **41**, 307–311 (2002).
38. Schreiber, S. L. Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science* **287**, 1964–1969 (2000).
39. Burke, M. D. & Schreiber, S. L. A planning strategy for diversity-oriented synthesis. *Angew. Chem. Int. Edn Engl.* **43**, 46–58 (2004).
40. Burke, M. D., Berger, E. M., Schreiber, S. L. Generating diverse skeletons of small molecules combinatorially. *Science* **302**, 613–618 (2004).
41. Walsh, C. *Antibiotics: Actions, Origins, Resistance* (ASM, Washington, 2003).
42. Hubbard, B., Walsh, C. Vancomycin assembly: Nature's way. *Angew. Chem. Int. Edn Engl.* **730**–765 (2003).
43. Schwarzer, D., Firking, R., Marahiel, M. A. Nonribosomal peptides: from genes to products. *Nat. Prod. Rep.* **20**, 275–287 (2003).
44. Walsh, C., Freil Meyers, C. L., Losey, H. C. Antibiotic glycosyltransferases: antibiotic maturation and prospects for reprogramming. *J. Med. Chem.* **46**, 3425–3436 (2003).
45. Wendt, K. U., Schulz, G. E., Corey, E. J., Liu, D. R. Enzyme mechanisms for polycyclic triterpene formation. *Angew. Chem. Int. Edn Engl.* **39**, 2812–2833 (2000).
46. Whittington, D. A. *et al.* Bornyl diphosphate synthase: structure and strategy for carbocation manipulation by a terpenoid cyclase. *Proc. Natl. Acad. Sci. USA* **99**, 15375–15380 (2002).
47. Rawlings, B. J. Type I polyketide biosynthesis in bacteria (part A). *Nat. Prod. Rep.* **18**, 190–230 (2001).
48. Rawlings, B. J. Type I polyketide biosynthesis in bacteria (part B). *Nat. Prod. Rep.* **18**, 231–281 (2001).
49. Kohli, R. & Walsh, C. Enzymology of acyl chain macrocyclization in natural product biosynthesis. *Chem. Commun.* 297–307 (2003).
50. Schwesche, T. *et al.* The biosynthetic gene cluster for the polyketide immunosuppressant rapamycin. *Proc. Natl. Acad. Sci. USA* **92**, 7839–7843 (1995).
51. Ahlert, J. *et al.* The calicheamycin gene cluster and its iterative type I enediynes PKS. *Science* **297**, 1173–1176 (2002).
52. Liu, W., Christensen, S. D., Standage, S. & Shen, B. Biosynthesis of the enediynes antitumor antibiotic C-1027. *Science* **297**, 1170–1173 (2002).
53. Shen, B., Liu, W. & Nonaka, K. Enediynes natural products: biosynthesis and prospects towards engineering novel antitumor agents. *Curr. Med. Chem.* **10**, 2317–2325 (2003).
54. Zerbe, K. *et al.* Crystal structure of OxyB, a cytochrome P450 implicated in an oxidative phenol coupling reaction during vancomycin biosynthesis. *J. Biol. Chem.* **277**, 47476–47485 (2002).
55. Pylypenko, O., Vitali, F., Zerbe, K., Robinson, J. A. & Schlichting, I. Crystal structure of OxyC, a cytochrome P450 implicated in an oxidative C–C coupling reaction during vancomycin biosynthesis. *J. Biol. Chem.* **278**, 46727–46733 (2003).
56. Walker, K. & Croteau, R. Taxol biosynthetic genes. *Phytochemistry* **58**, 1–7 (2001).
57. Chau, M., Jennewein, S., Walker, K. & Croteau, R. Taxol biosynthesis: molecular cloning and characterization of a cytochrome P450 taxoid 7 beta-hydroxylase. *Chem. Biol.* **11**, 663–672 (2004).
58. Jennewein, S., Long, R. M., Williams, R. M. & Croteau, R. Cytochrome p450 taxadiene 5 alpha-hydroxylase, a mechanistically unusual monooxygenase catalyzing the first oxygenation step of taxol biosynthesis. *Chem. Biol.* **11**, 379–387 (2004).
59. Yadav, G., Ghokale, R. S. & Mohanty, D. Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases. *J. Mol. Biol.* **328**, 335–363 (2003).
60. Eppelman, K., Stachelhaus, T. & Marahiel, M. A. Exploitation of the selectivity-conferring code of nonribosomal peptide synthetases for the rational design of novel peptide antibiotics. *Biochemistry* **42**, 9718–9716 (2002).
61. Lu, W., Leimkuhler, C., Oberthur, M., Kahne, D. & Walsh, C. AklN is an L-2-deoxyfucosyltransferase in the biosynthesis of the anthracycline aclacinomycin. *Biochemistry* **43**, 4548–4558 (2004).
62. Losey, H. C. *et al.* Incorporation of glucose analogs by glycosyltransferases GtfE and GtfD from the vancomycin biosynthetic pathway to generate variant glycopeptides. *Chem. Biol.* **9**, 1305–1314 (2002).
63. MendeZ. C. & Salas, J. A. Altering the glycosylation pattern of bioactive compounds. *Trends Biotechnol.* **19**, 449–456 (2001).
64. McDaniel, R. *et al.* Multiple genetic modifications of the erythromycin gene cluster to produce a library of novel 'unnatural' natural products. *Proc. Natl. Acad. Sci. USA* **96**, 1846–1851 (1999).
65. Tang, L. & McDaniel, R. Construction of desosamine containing polyketide libraries using a glycosyltransferase with broad substrate specificity. *Chem. Biol.* **8**, 547–555 (2001).
66. Wohlleben, W. & Pelzer, S. New compounds by combining modern genomics and old-fashioned mutasynthesis. *Chem. Biol.* **9**, 1163–1166 (2002).
67. Huang, Q., Roessner, C. A., Croteau, R. & Scott, A. I. Engineering *E. coli* for the synthesis of taxadiene, a key intermediate in the biosynthesis of taxol. *Biorg. Med. Chem.* **9**, 2237–2242 (2001).
68. Martin, V. J., Pitera, D. J., Withers, S. T., Newman, J. D. & Keasling, J. D. Engineering a mevalonate pathway in *E. coli* for production of terpenoids. *Nature Biotechnol.* **21**, 796–802 (2003).
69. Ondeyka, J. G. *et al.* Nodulisporic acids C₁, C₂, and C₃: a series of D-ring-opened nodulisporic acids from the fungus *Nodulisporium* sp. *J. Nat. Prod.* **66**, 121–124 (2003).
70. Brady, S. F., Bondi, S. M. & Clardy, J. The guanacastepenes: a highly diverse family of secondary metabolites produced by an endophytic fungus. *J. Am. Chem. Soc.* **123**, 9900–9901 (2001).
71. Fuller, R. W. *et al.* A pentahalogenated monoterpene from the red alga *Portieria hornemanni* produces a novel cytotoxicity profile against a diverse panel of human tumor cell lines. *J. Med. Chem.* **35**, 3007–3011 (1992).
72. Piel, J. Metabolites from symbiotic bacteria. *Nat. Prod. Rep.* **21**, 519–538 (2004).
73. Luesch, H., Yoshida, W. Y., Moore, R. E., Paul, V. J. & Corbett, T. H. Total structure determination of apratoxin A, a potent novel cytotoxin from the marine cyanobacterium *Lyngbya majuscula*. *J. Am. Chem. Soc.* **123**, 5418–5423 (2001).
74. Edwards, D. J. *et al.* Structure and biosynthesis of the jamaicamides, new mixed polyketide-peptide neurotoxins from the marine cyanobacterium *Lyngbya majuscula*. *Chem. Biol.* **11**, 817–833 (2004).
75. Gerth, K., Bedorf, N., Hofle, G., Irschik, H. & Reichenbach, H. Epithilons A and B: antifungal and cytotoxic compounds from *Sorangium cellulosum* (Myxobacteria). Production, physico-chemical and biological properties. *J. Antibiot. (Tokyo)* **49**, 560–563 (1996).
76. Sasse, F., Steinmetz, H., Heil, J., Hofle, G. & Reichenbach, H. Tubulysins, new cytostatic peptides from myxobacteria acting on microtubuli. Production, isolation, physico-chemical and biological properties. *J. Antibiot. (Tokyo)* **53**, 879–885 (2000).
77. Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740 (1997).
78. Kaerberlein, T., Lewis, K. & Epstein, S. S. Isolating 'uncultivable' microorganisms in pure culture in a simulated natural environment. *Science* **296**, 1127–1129 (2002).
79. Wang, G. Y. *et al.* Novel natural products from soil DNA libraries in a streptomycete host. *Org. Lett.* **2**, 2401–2404 (2000).
80. Brady, S., Chao, C. J. & Clardy, J. New natural product families from an cDNA cluster. *J. Am. Chem. Soc.* **124**, 9968–9969 (2002).
81. Zazopoulos, E. *et al.* A genomics-guided approach for discovering and expressing cryptic metabolic pathways. *Nature Biotechnol.* **21**, 187–190 (2003).
82. Jennewein, S., Wildung, M. R., Chau, M., Walker, K. & Croteau, R. Random sequencing of an induced *Taxus* cell cDNA library for identification of clones involved in Taxol biosynthesis. *Proc. Natl. Acad. Sci. USA* **101**, 9149–9154 (2004).

Acknowledgements We thank G. J. Gatto and K. N. Maloney for their help in preparing this article.

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Natural and engineered nucleic acids as tools to explore biology

Ronald R. Breaker

Department of Molecular, Cellular and Developmental Biology, Yale University, P. O. Box 208103, New Haven, Connecticut 06520-8103, USA
(e-mail: ronald.breaker@yale.edu)

RNA and DNA molecules can form complex, three-dimensional folded structures that have surprisingly sophisticated functions, including catalysing chemical reactions and controlling gene expression. Although natural nucleic acids make occasional use of these advanced functions, the true potential for sophisticated function by these biological polymers is far greater. An important challenge for biochemists is to take RNA and DNA beyond their proven use as polymers that form double-helical structures. Molecular engineers are beginning to harness the power of nucleic acids that form more complex three-dimensional structures, and apply them as tools for exploring biological systems and as therapeutics.

Exploring the full complexity of cells at the molecular level will require the fashioning of new tools that allow researchers to manipulate complex biological processes in unique ways. Small organic molecules that block or otherwise perturb the normal functions of the cellular machinery have long served as powerful tools for exploring biochemical processes. Similarly, new tools that take advantage of the natural functions of proteins and nucleic acids are proving to be enormously useful as researchers continue to probe the details of complex biochemical systems.

Living systems have been expanding and diversifying their natural collection of biochemical tools for billions of years. For example, enzymes build RNA, DNA and proteins with high fidelity and with impressive speed; in some cases more than 100 monomeric units are added to the polymer per second. Many other enzymes are known to selectively cut or join nucleic acids or proteins, and still others catalyse chemical reactions with great speed and accuracy. This provides us with a large set of verified technologies which, if harnessed by researchers, can be applied to understand and manipulate biological processes at their most fundamental level. Indeed, there is a considerable history of scientists taking bits and pieces of proteins and nucleic acids from natural sources, tailoring them by purposefully mutating or splicing them in different ways, and using them as reagents for biological study or for therapeutic applications.

More recently, researchers have begun to harness darwinian evolution to optimize existing functions of proteins^{1,2} and nucleic acids^{3,4}, and to create new ones. In combination with rational design methods, these techniques for directing the evolution of biopolymers allow researchers to become a creative force for molecular change and invention. In many instances, we no longer need to be limited to using a less-than-optimal protein or nucleic acid molecule from natural sources. Some natural proteins and nucleic acids can be enhanced by using directed evolution or entirely new functions can be derived using similar engineering strategies.

Simple, engineered nucleic acids already provide us with useful tools for detecting and manipulating other nucleic acids. For example, the selective amplification of genomic fragments by the polymerase chain reaction (PCR)⁵ or by related techniques requires the use of designed synthetic DNA primers. Similarly, the targeted inactivation of gene expression by using short synthetic oligonucleotides or small interfering RNAs (siRNAs)^{6,7} is becoming increasingly routine. These applications are greatly aided by efficient

methods for the sequence-specific chemical^{8,9} and enzymatic¹⁰ synthesis of RNA and DNA. In addition, the design of nucleic acids that bind to other nucleic acids with high affinity and specificity follows the simple and long-established rules of Watson–Crick base pairing¹¹.

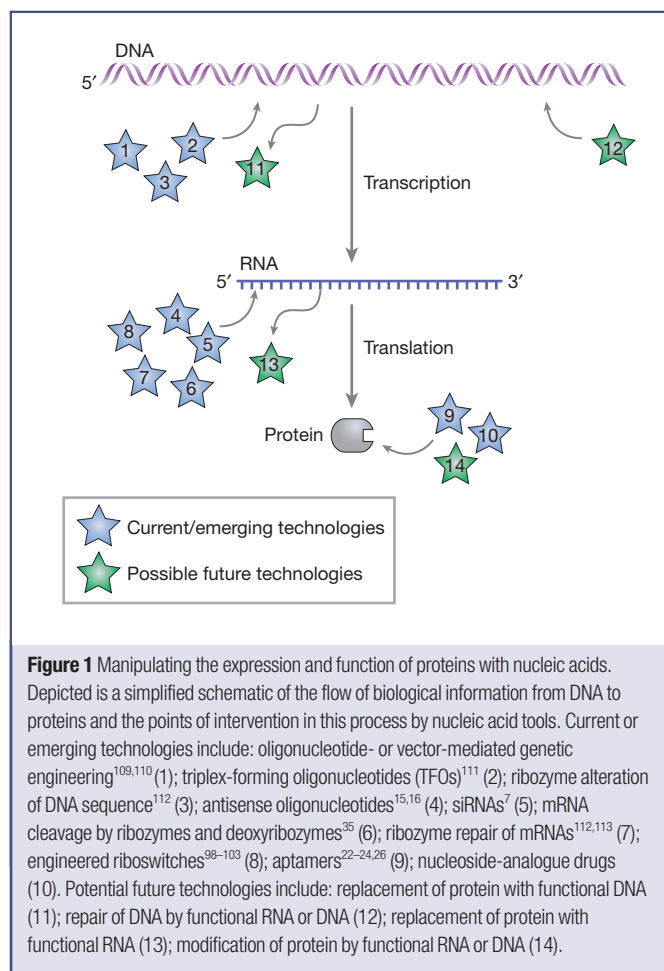
However, it is becoming increasingly clear that nucleic acids can have far greater use than that shown by simple base-paired structures. For example, the hammerhead ribozyme consists of just over 30 nucleotides and can catalyse RNA-strand scission at a rate that is millions of times faster than spontaneous RNA cleavage¹². At the opposite end of the spectrum is the ribosome, which at its core carries a staggeringly complex ribozyme structure that catalyses peptide-bond formation^{13,14} (see section ‘Ribozymes and deoxyribozymes’ below).

These natural RNAs are just a small representation of the considerable untapped potential that nucleic acids have for forming complex structures and carrying out sophisticated tasks; it is this potential that nucleic acid engineers seek to harness. Nucleic acid design and synthesis techniques, along with powerful directed evolution strategies, are empowering the drive to design ever more complex RNA and DNA molecules. Here, I will focus on some of the surprising functions of novel ‘designer’ nucleic acids, and assess the potential for these new tools in biotechnology and therapeutics.

Manipulating life's central processes

To gauge the potential use of engineered nucleic acids in manipulating biological systems, we need only consider the roles of DNA and RNA in fundamental biological processes. The basic roles for DNA and RNA in information storage and transfer are well established, and numerous ways in which this process can be manipulated by using nucleic acids are being explored. Many existing technologies and several emerging ones can be used to selectively target gene expression and protein function at the DNA, RNA and protein levels (Fig. 1).

The instructions for protein synthesis, encoded by the nucleotide sequences of genomic DNA, are transferred to messenger RNAs that are subsequently ‘read’ by ribosomal RNAs and transfer RNAs. Therefore, the information stored in DNA or RNA can be manipulated by designing short complementary DNA or RNA oligonucleotides that bind the nucleic acids. For example, antisense oligonucleotides (whose sequences are complementary to their target genes) have been developed to selectively inhibit a variety of



genes^{15,16}; one such molecule (Vitravene)¹⁷ has been commercialized as an antiviral agent¹⁸. So, as with siRNAs, designing new oligonucleotides that downregulate gene expression can be as simple as creating a complementary sequence for the target mRNA.

However, antisense molecules typically function by different mechanisms¹⁵ from those used by siRNAs¹⁹, and thus do not take advantage of the natural siRNA processing enzymes. In general, more research and development is needed to ensure that each antisense oligonucleotide effectively targets its intended mRNA. Moreover, these and other approaches that use oligonucleotides *in vivo* must ensure that the RNA or DNA molecules being delivered are sufficiently resistant to chemical and enzymatic degradation in a cellular environment. Already, there have been numerous advances in the chemical synthesis of nucleic acid analogues²⁰. These allow the oligonucleotide polymers to persist in the bloodstream for many hours, where otherwise they would have a half-life of seconds.

Another major role for nucleic acids in fundamental biological processes is not as polymers but as nucleotide-like fragments of essential metabolites and coenzymes. Small nucleotides like ATP and GTP, the cyclic nucleotides cAMP and cGMP, and numerous coenzymes and metabolic intermediates, including nucleotide fragments, are involved in many metabolic and signalling pathways. The proteins in these pathways are exploited by many drug compounds that mimic the basic structures of nucleotides and nucleotide-like coenzymes. Similarly, various nucleoside analogues (such as human immunodeficiency virus (HIV) reverse transcriptase chain terminators)²¹ show antiviral activity because they interfere directly with the synthesis of new pathogen DNAs.

The purpose of this review, however, is to discuss the tremendous potential for more complex, folded nucleic acids which carry out functions that, until recently, had only been observed in proteins.

These functions result when RNAs and DNAs form more globular structures, which usually include both helical structures and long-distance tertiary contacts, such as atypical molecular contacts between nucleotides and metal-ion binding to bases and phosphates. This means that researchers are not limited to exploiting the rules of Watson–Crick base pairing or designing nucleotide-like compounds that fortuitously occupy binding sites on proteins.

The diversity of sophisticated functions undertaken by structured nucleic acids opens many opportunities to create new tools that can be used to explore biological systems. Hints of this potential have been emerging in recent years with the development of aptamers (ligand-binding polynucleotides), ribozymes, deoxyribozymes and riboswitches (metabolite-sensing gene control elements). As with proteins, functional nucleic acids can be isolated from natural sources. But to tap into the full potential for structured nucleic acids, researchers will need to use engineering strategies, such as directed evolution^{3,4}, which can be used to generate RNAs and DNAs with entirely new tertiary structures. Specifically, this process works by selectively reproducing copies of RNAs or DNAs that have performed some task, such as ligand binding or self-cleavage. Some researchers believe that if an experimental protocol can be devised to identify functional molecules from trillions of inactive variants, then it is likely that a nucleic acid will be found that performs the desired task (as long as the task is compatible with the principles of chemistry). Some of the key advances in using directed evolution to acquire nucleic acid tools are described in the next section.

RNA and DNA aptamers

Dramatic examples of nucleic acids performing more complex functions have been provided by researchers who create and study aptamers (Fig. 2). Engineered aptamers are structured RNA or DNA molecules that form binding pockets for specific ligands^{22,23}. They can be created by using directed evolution techniques: trillions of RNA or DNA molecules are prepared simultaneously and subjected to a process of selective amplification to enrich the population with variants that bind to a particular protein target. Directed or *in vitro* evolution of aptamers requires that some distinction be made between RNAs or DNAs that bind to a target ligand and those that do not. This is usually achieved by using some form of affinity chromatography to physically separate ligand-binding polynucleotides from the vast population of inactive variants. Isolation is then followed by amplification of the rare molecules.

Alternatively, domains of random sequences can be integrated with catalytic RNAs or DNAs such that ligand binding to specific variants triggers catalytic function. This arrangement, wherein ligand binding at one site controls the activity of a distant catalytic site, is defined as allosteric enzyme activity (see section on 'Allosteric ribozymes' below). For example, self-cleaving ribozymes that are triggered to cleave only when incubated in the presence of cyclic nucleotide monophosphates (for example, cAMP) have been created by using 'allosteric selection'²⁴. Cleaved RNAs are then physically separated from those that are not cleaved, permitting the experimenter to isolate ligand-binding RNAs from large pools of random sequences. These composite aptamer–ribozyme constructs can subsequently be deconstructed to yield separate aptamer domains that retain their ligand-binding function²⁵.

Many aptamers have functional characteristics that are similar to antibodies²⁶. Just like protein antibodies, aptamers can selectively recognize specific protein or small-molecule ligands, even in complex chemical or biological mixtures. Furthermore, they can bind to their cognate targets at target concentrations of the nanomolar or picomolar range, matching or even superseding the affinities of antibodies. Aptamers retain their function when they are immobilized for use in the test tube; they can also be delivered to organisms or expressed inside cells. Aptamers generated in the test tube can be made to bind a diverse array of targets, including highly toxic agents^{27,28}, and may perform their receptor functions under defined conditions that are far

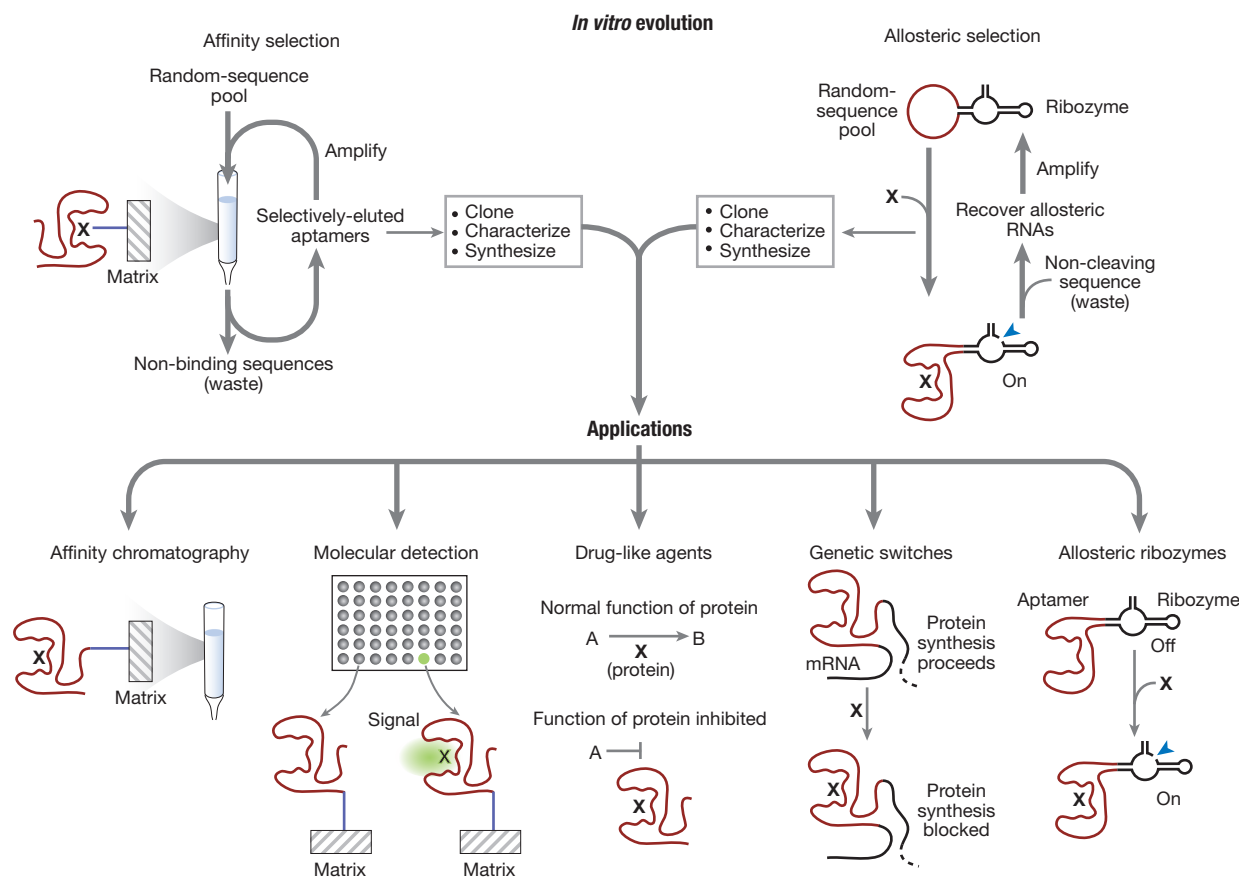


Figure 2 The generation and application of aptamers. Methods for the *in vitro* evolution of aptamers include the use of an affinity matrix to separate RNA variants that selectively bind an immobilized target (X; X represents any target molecule unless otherwise noted)^{22,23,26}, or the use of allosteric ribozymes that permit

separation of ligand-binding RNAs by means of self-cleavage^{81–83}. Once they have been engineered, aptamers can be used (among other applications) as chromatographic agents, biosensor elements, anti-protein drugs, gene-control elements and as components of allosteric ribozymes.

from the physiological norm. This last feature might be useful for biosensing applications, for which solvent conditions or desired temperatures might be disruptive to protein structures.

As with antibodies, the potential use of aptamers is considerable (Fig. 2). Aptamers can be immobilized on solid supports to yield designer matrices for affinity chromatography^{29,30}. In most instances, aptamers undergo a change in shape on ligand binding³¹. This can be exploited to create biosensors (devices that use biological materials to monitor the presence of various chemicals in a substance) by the judicious integration of fluorescent tags with aptamers in solution or immobilized on surfaces. Such aptamer beacons have been used to detect specific proteins^{32,33} or small molecules³⁴ through the change in fluorescence that occurs upon ligand binding.

Perhaps most profound is the possible application of protein-binding aptamers as therapeutic agents. A tremendous body of literature exists regarding the prospects for using nucleic acids for *in vivo* applications³⁵. But, as is the case for other oligonucleotides, the delivery of aptamers into cells is problematic³⁶. Also, the molecules must be resistant to nuclease degradation or they risk being rapidly destroyed by nucleases in the blood. Various approaches are being developed to improve the capabilities of aptamers *in vivo*. For example, the isolation of new aptamers (and ribozymes) can be conducted with chemical modifications already in place to facilitate nuclease resistance of the winning molecules, such as phosphorothioate linkages or modifications at the 2' carbon of ribose^{37,38}. A

number of examples now exist in which aptamers expressed inside cells retain their function and bring about their intended inhibitory effect. These aptamers, sometimes called 'intramers'^{39,40}, can be stabilized by integrating them into a larger RNA construct. This RNA construct is inherently more resistant to degradation or can help route the RNA to its desired cellular compartment⁴¹.

Furthermore, mirror-image aptamers, called *spiegelmers*^{42,43}, can be created to serve as highly stable receptors for their corresponding ligands. Spiegelmers have a chiral configuration (L-RNA) which is the mirror image of that for 'normal' or D-RNA. Despite this apparently radical alteration, *spiegelmers* can be made by using *in vitro* evolution in much the same way as that used to generate normal aptamers. First, normal D-RNA aptamers are generated that bind to the mirror image of the target that one wishes to bind with a *spiegelmer*. For example, an unnatural peptide target that is the mirror image (D-polypeptide) of that normally encountered in the cell (L-polypeptide) is used during the selection and amplification process. So, the resulting normal D-RNA aptamer would be functionally useless against the natural target analogue. Once in hand, however, the sequence of the normal D-RNA aptamer serves as a guide to make the L-configured *spiegelmer* by using L-nucleotides during chemical synthesis. This simple production trick creates an L-aptamer that can bind an L-polypeptide, starting from the D-aptamer–D-polypeptide complex that was originally created by *in vitro* evolution. Although spontaneous degradation by inherent

chemical instability of RNA should remain unchanged, spiegelmers are completely resistant to degradation by typical nucleases⁴⁴.

One of the challenging aspects of this technology is that new aptamers are not always easy to generate. The basic protocols for *in vitro* evolution are rather straightforward, but trivial problems with any of the selection or amplification steps can cause bottlenecks that restrict molecular variation, or can cause complete loss of the evolving population. Anything from losing trace amounts of nucleic acids because of non-specific binding to plastic tubes, to technical problems that create DNA amplification artifacts during PCR, can sap the efficiency of *in vitro* evolution experiments. Equally problematic is the emergence of 'selfish' RNA or DNA molecules. These 'molecular weeds' typically lack the desired ligand-binding function, but use alternative strategies to survive the selection process. For example, many aptamers have been isolated that bind to the chromatographic matrix (agarose, nitrocellulose) as opposed to the ligand that is immobilized on the matrix. Precautions can be taken to avoid or eliminate problems encountered during *in vitro* evolution, but these require additional steps or judgements to be made at each stage of the selection process. For example, matrix-binding aptamers can be disfavoured by using free ligand to selectively recover the desired aptamers from the chromatographic matrix. Free ligands compete for the aptamer binding sites, causing selective elution of aptamers that are bound to immobilized ligand versus those that simply bind the matrix.

Until recently, aptamer generation was a completely manual operation: it involved numerous pipetting and purification steps that had to be conducted with great care. However, there are several reports^{45–48} of successful automated aptamer selections that require minimal hands-on effort. Aptamers produced by automated methods target diverse proteins, such as lysozyme⁴⁵ and the human U1A protein⁴⁸. These automated methods and other manual protocols permit the pursuit of aptamers for many targets simultaneously. So far, promising drug-like aptamers, such as an anti-VEGF (vascular endothelial growth factor) aptamer⁴⁹ and two anti-clotting aptamers^{50,51}, have been created using manual selection methods. If hundreds or thousands of aptamers are demanded, technology appears to be advancing to the point where they could in principle be generated.

Ribozymes and deoxyribozymes

To date, there are nine known classes of natural ribozymes that catalyse phosphoester cleaving/forming or peptide-bond-forming reactions. The peptide-bond forming ribozyme is found at the core of the ribosome and is made up of the most highly conserved segments of rRNA^{14,52}. Although this ribozyme has long been the target of many antibiotic drugs^{53–55}, there is considerable interest in using ribozymes themselves as therapeutic agents⁵⁶. For example, RNase P, a phosphodiester-cleaving ribozyme that normally processes tRNA precursors, can be induced to cleave new RNA targets (such as mRNAs): an external 'guide' sequence that is delivered or expressed in cells directs the cleavage event^{57–59}. A mix of conventional base pairing and tertiary structure formed by the guide sequence when docked to its target RNA is recognized as a substrate to be cleaved by RNase P. Two types of self-splicing ribozyme, called group I and group II because of their distinct structures and reaction mechanisms, have also been designed to catalyse *trans*-splicing of mRNAs^{60,61} or to direct their own integration into genomic DNAs to yield genetic changes⁶². If these ribozymes were made to efficiently modify the mRNAs or the DNAs that serve as their templates, they could be used as new gene-repair systems.

Other ribozymes show promise as agents for the destruction of viral RNA or mRNAs. Indeed, most efforts to make therapeutic ribozymes have been directed towards developing the small self-cleaving ribozymes³⁵ into selective mRNA-cleaving agents. For example, the hammerhead ribozyme (Fig. 3) can be made to cleave different RNA substrates simply by tailoring the nucleotide sequences of its substrate-binding arms. RNA-cleaving ribozymes

configured for therapeutic applications or for target validation⁶³ typically catalyse one reaction per minute. This is fast enough to have a biological impact, but not sufficiently fast to permit each ribozyme to process more than just a few substrate molecules before it is destroyed by cellular enzymes.

In vitro evolution can be used to create new ribozymes that catalyse RNA cleavage as well as many other chemical reactions^{3,4}. Although there might not be an immediate use for self-alkylating RNAs⁶⁴ or for ribozymes that form the glycosidic linkage of nucleotides⁶⁵, other engineered ribozymes might find application much sooner. For example, ribozymes can be made to covalently attach to specific proteins⁶⁶, suggesting that designer ribozymes could be created that selectively couple to many different potential therapeutic or diagnostic protein targets.

A ribozyme-catalysed reaction that has more obvious use is that of RNA cleavage. The diversity of motifs that catalyse RNA cleavage by internal phosphoester transfer is substantial. Therefore, it is possible to design new sequences that have reaction characteristics tuned to the desired application. Some of these engineered ribozymes, such as the X-motif^{67,68} (Fig. 3), have performance characteristics that are similar to that of the hammerhead ribozyme, indicating that new ribozymes could be created that destroy disease-causing RNAs with an efficiency equal to or greater than natural ribozymes.

Even DNA enzymes or 'deoxyribozymes' have been created that cleave RNA by using the same phosphoester transfer reaction⁶⁹. The most studied catalytic DNA is the 10–23 deoxyribozyme⁷⁰ (Fig. 3). As with the hammerhead and X-motif ribozymes, 10–23 can be tailored to cleave almost any RNA molecule, and its catalytic efficiency allows it to affect gene expression⁷¹. For example, chemically synthesized 10–23 has been used to reduce the expression of a gene responsible for undesired tissue growth after artery damage^{72,73}. As with other oligonucleotide therapeutics, there are concerns about deoxyribozyme delivery and pharmacokinetics, DNA stability, subcellular localization and biochemical access to target sites. However, results to date demonstrate that both ribozymes and deoxyribozymes can

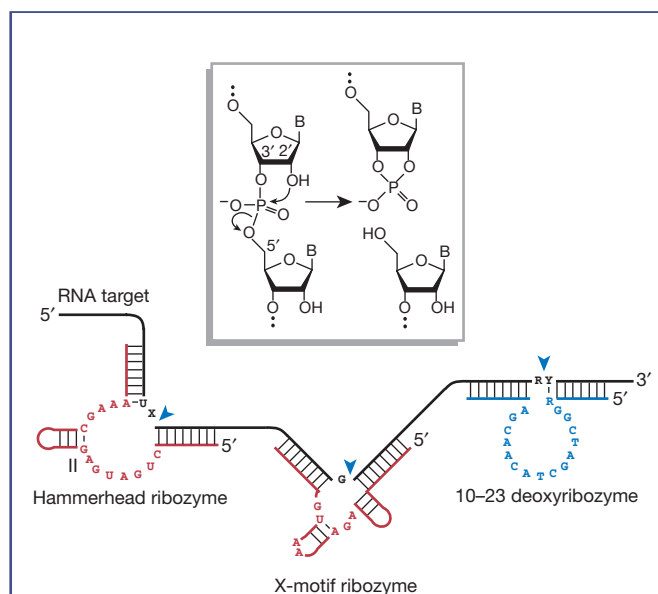


Figure 3 RNA-cleaving ribozymes and deoxyribozymes. The natural hammerhead ribozyme as well as the engineered X-motif ribozyme and 10–23 deoxyribozyme motifs catalyse RNA cleavage by promoting an internal phosphoester transfer reaction (inset). Base pairing between the RNA target and the substrate-binding arms of each catalyst can be tailored to target different RNA sequences. Nucleotides within the target RNA, the ribozymes and the deoxyribozyme that are not conserved are depicted with black, red and blue lines, respectively. B, base.

indeed function as agents for downregulating gene expression in a targeted fashion.

As noted above, engineered ribozymes and deoxyribozymes can catalyse reactions other than RNA cleavage. Some of these reactions would be ideal for manipulating the chemical structures of proteins and nucleic acids, either *in vitro* or *in vivo*. For example, ribozymes that selectively ligate RNA to protein have been created by *in vitro* evolution⁶⁶. If these ribozymes were made to function inside cells, it might be possible to manipulate protein function by new mechanisms. In addition, numerous deoxyribozymes that use ATP to phosphorylate DNA^{74,75} or to ligate DNA⁷⁶ have been generated. Although the efficiencies of most of these deoxyribozymes are far from sufficient to be of biological relevance, improvements to their action could be made. If so, unique tools for manipulating biological polymers would result (Fig. 1).

Allosteric ribozymes

The simplicity of the interactions that define the secondary structures of RNA and DNA molecules causes complications for those who study the structures and functions of nucleic acids. As the length of the molecule increases, so do the number of opportunities to form alternative base-pair or tertiary-structure interactions that prevent the desired fold from forming. Many different folding pathways that yield many alternately folded (and inactive) states are possible⁷⁷. If

conditions are right and the alternately folded structures are not very stable, these states can interchange on a timescale that is sufficiently short for the interchangeable state to be harnessed for useful functions. So, the conformational heterogeneity of nucleic acids can be turned into an important positive characteristic: this has been achieved by molecular engineers and in spectacular fashion by natural mRNAs.

For example, an ATP-binding aptamer created by *in vitro* evolution carries two base-paired elements that are pre-formed in the absence of ligand, whereas its ligand-binding core remains largely disordered^{78,79} (Fig. 4a). However, the docking of ATP stabilizes the aptamer's core and lends additional stability to the adjoining base-paired stems. When it is appropriately fused to a weakened but essential stem of a hammerhead ribozyme, the aptamer acts as an allosteric binding site and permits the ribozyme to be activated by ATP binding⁸⁰.

This simple demonstration of allosteric activation of ribozymes has since been expanded upon; numerous RNA switches that are selectively triggered by many signals (including small organic compounds, proteins, nucleic acids, metal ions, pH and light) have been created^{81,82}. Each engineered RNA switch can independently serve as a biosensor element for its corresponding ligand. Immobilized RNAs that are tagged with radioactive or fluorescent labels have been used to form biosensor arrays that report the presence and concentrations of targets, even in complex chemical or biological mixtures^{83,84}. In other manifestations, ribozymes have been shown to function as diagnostics that sense the presence of pathogen-specific molecules, such as viral RNAs^{85,86}.

In these examples, the RNAs are not replacing the action of a small molecule but they could ultimately be used to identify small molecules that affect biological functions. For example, protein kinases typically convert ATP into ADP upon protein phosphorylation. A highly specific ADP-sensing RNA switch or RiboReporter^{87,88} has been created and used to detect and report the amount of ADP by-product (Fig. 4b). This indirectly reflects the level of protein kinase activity in a given assay. A RiboReporter that yields a fluorescent report upon activation by ADP was used successfully to identify reaction mixtures containing the protein kinase inhibitor staurosporine (Fig. 4c)⁸⁷. Similarly, this allosteric ribozyme could be used in large high-throughput screens to identify new protein kinase inhibitors or to find compounds that modulate the activity of any enzyme whose activity generates or destroys ADP. Allosteric ribozymes have also been created to respond to protein targets⁸⁹ such as lysozyme and the Rev peptide from HIV⁸⁴. Furthermore, other protein-dependent allosteric ribozymes have recently been shown to be useful tools to screen for small molecules that disrupt protein–ligand interactions⁹⁰.

Riboswitches

Recent studies have begun to reveal that many bacteria already make extensive use of natural RNA aptamers for metabolite sensing and gene-control purposes^{91,92}. These natural RNA switches, or riboswitches^{93,94}, show a wide range of target specificities and affinities. For example, they are known to be responsible for controlling the expression of about 2% of the genes in *Bacillus subtilis*⁹⁵. Not only does the existence of riboswitches add validity to the notion that useful RNA switches can be engineered, but this mechanism for gene control also offers numerous opportunities to use natural or engineered aptamers *in vivo* for new applications.

The adenine-responsive riboswitch^{95,96} from *B. subtilis* has characteristics that are typical of most other riboswitches. The riboswitch carries an aptamer domain that conforms to a consensus sequence and secondary structure (Fig. 5a). The aptamer resides immediately upstream of an 'expression platform' that modulates gene expression in response to metabolite binding. Its function is similar to that of the linker regions between aptamers and ribozymes in engineered RNA switches. Of the several mechanisms used by riboswitches in *B. subtilis*, the regulation of transcription through metabolite-mediated control of transcription termination is the most common. The adenine-specific

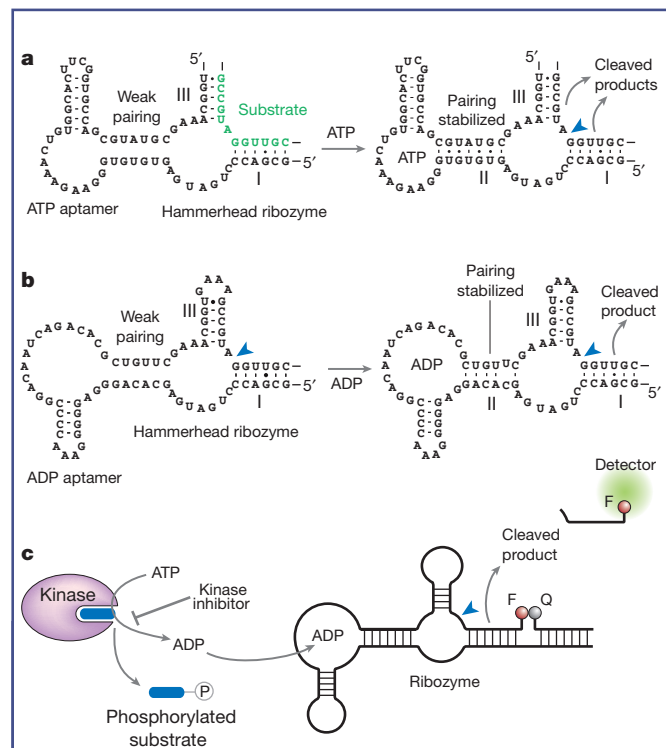


Figure 4 Allosteric ribozymes as precision biosensor elements. **a**, One of the first engineered allosteric ribozymes was created by fusing an ATP-binding aptamer to a hammerhead ribozyme by means of a disordered bridge element⁸⁰. Ligand binding stabilizes the core of the aptamer and the weakly pairing stem (stem II of the ribozyme) to trigger increased ribozyme activity. **b**, A next-generation allosteric ribozyme or RiboReporter that senses ADP and disfavours binding of ATP by more than 100-fold. **c**, In the design shown here, a fluorescent readout is generated if ribozyme activity is triggered by ADP. As a result, fluorescence increase is prevented if an anti-protein-kinase drug, such as staurosporine, is present⁸⁷. RNA cleavage by the ribozyme occurs within the stem I/III junction between A and G (blue arrow). F and Q represent fluorophore and quencher moieties, respectively. The performance characteristics of this RNA switch are sufficient to permit its use in high-throughput screening assays.

riboswitch shown in Fig. 5a uses this mechanism to activate expression of a gene encoding an adenine efflux pump when excess adenine is present^{96,97}. Similar riboswitches repress gene expression upon introduction of the target metabolite. Given that these RNAs are highly modular and can be moved from one gene to the next, there is considerable potential to create transgenic organisms that express genes in response to several different metabolites.

Even more useful would be the creation of designer riboswitches that have entirely new ligand specificities. Already, several studies

report the successful integration of aptamers with mRNAs to permit ligand-specific gene control^{98–103}. For example, a reporter gene construct was made to express a theophylline-binding aptamer (itself previously created using *in vitro* evolution) located immediately upstream of its ribosome-binding site (RBS) for the coding region of the reporter-gene mRNA¹⁰³. This aptamer–mRNA fusion allowed gene expression to be controlled by the addition of theophylline to a bacterial cell culture (Fig. 5b). Further iterations of these engineering efforts, perhaps augmented by reverse engineering of natural riboswitches, promise to provide designer gene-control switches for a variety of applications, such as *in vivo* metabolite sensing and/or the control of therapeutic genes delivered by retroviral vectors. It is already known that certain eukaryotic cells carry riboswitches¹⁰⁴. Therefore, it seems reasonable to speculate that engineered riboswitches could function as designer gene-control elements in humans without provoking an undesired immune response as occurs with protein-based systems.

The ability to create new aptamers and riboswitches offers a way to create functions for non-natural compounds in gene control. It is also interesting to note that, because riboswitches have evolved to purposefully bind to metabolites, they should be able to serve as targets for drug compounds — much like their protein receptor counterparts. Indeed, it is now clear that riboswitches already serve as drug targets. The compound aminoethylcysteine (AEC), which for many years has been known to be toxic to bacterial cells, appears to work at least in part by binding to lysine-specific riboswitches and causing downregulation of lysine biosynthetic genes¹⁰⁵. It is therefore likely that additional small compounds could be created to serve as anti-infective agents by targeting other crucial bacterial riboswitches.

Conclusions

The most mature nucleic acid technologies, such as those using anti-sense RNA and siRNAs, are widely used to selectively knock out the function of certain proteins by inhibiting gene expression. Engineered aptamers and designer enzymes can be used to modulate protein action once the polypeptide has been made; or these aptamers and enzymes can serve as biocatalysts in their own right. A high level of validation for these technologies exists in nature. For example, the most recently discovered class of ribozyme was found to be encoded upstream of the *glmS* gene in *B. subtilis*¹⁰⁶. The *glmS* ribozyme is a small self-cleaving RNA that surprisingly also functions as a metabolite-sensing molecular switch. The ribozyme is selectively activated by a natural sugar compound, glucosamine-6-phosphate; its activity downregulates the expression of the protein that produces the sugar metabolite. This highlights the potential use of aptamers and ribozymes as agents for genetic control.

Other technologies, such as ribozyme-mediated modification of proteins or the complete replacement of a protein with an engineered nucleic acid, will require a substantial amount of research and development to make practical. However, the functional diversity of nucleic acids is enormous and future engineering efforts will certainly be made to expand the function and use of RNA and DNA tools. Currently, most research and development in nucleic acid engineering is being carried out as part of the basic research programmes of individual laboratories in academia. These efforts continue to provide proof of principle for an ever-widening array of nucleic acid tools.

Unfortunately, however, university-based research laboratories are not always the most appropriate settings for the maturation of emerging biotechnologies. Thus, it is encouraging to note that several small biotechnology companies are pursuing some of the more immediate applications of engineered nucleic acids, such as therapeutic aptamers and engineered RNA genetic switches. Undoubtedly, the continued exploration of the functional potential of nucleic acids will create new opportunities for those who seek to harness engineered RNAs and DNAs for practical applications. This process could be accelerated by providing support for research and development projects aimed at creating simple tools that could become useful for

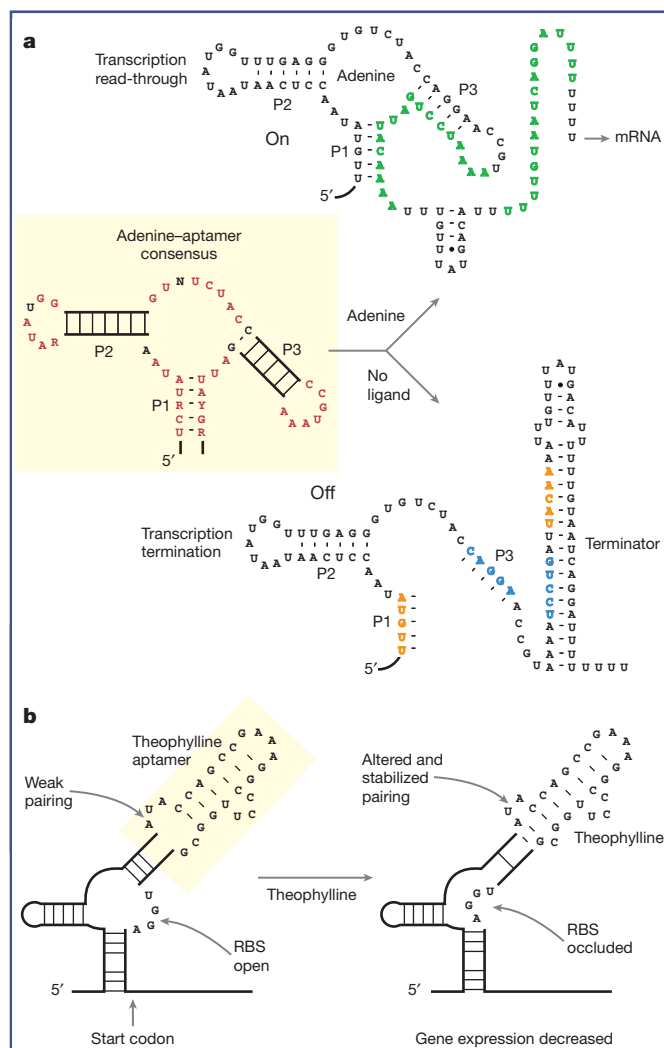


Figure 5 Natural and engineered riboswitches for controlling gene expression. **a**, A natural adenine-binding aptamer and its role in activating gene expression as part of an adenine riboswitch from the *ydhL* gene of *Bacillus subtilis*. The consensus sequence and secondary structure for the adenine aptamer domain is shaded. When sufficient adenine is present (top), the 5' untranslated region (UTR) folds to form the full secondary structure for the aptamer to bind to its ligand. This precludes nucleotides from forming an intrinsic terminator stem^{114,115} (shown in green) and a complete mRNA is synthesized. In the absence of adenine (bottom), portions of the secondary structure required for the aptamer to bind to its ligand (red and blue nucleotides) are not formed, which permits the intrinsic terminator to form and cause premature transcription termination. Gene expression is prevented because the complete mRNA is not synthesized. **b**, Proposed mechanism for an engineered genetic switch that uses a theophylline-specific aptamer¹⁰³. The aptamer (shaded) and a short linker region is integrated with a 5' UTR and the construct is fused upstream of an open reading frame. In the absence of theophylline (left), the protein is expressed because the RBS is available for interaction with ribosomes. In the presence of theophylline (right), the aptamer/linker structure becomes stabilized. This more stable structure presumably restricts ribosome access to the RBS, thus reducing gene expression.

basic research. Striking examples of this include the development of RNA-cleaving ribozyme constructs which, when expressed in cells, can be used to identify genes that are critical for certain cellular pathways^{107,108}. Some of these tools for basic research will probably develop into treatments and become new classes of drugs to complement traditional small molecules. □

doi:10.1038/nature03195

1. Kolkman, J. A. & Stemmer, P. C. Directed evolution of proteins by exon shuffling. *Nature Biotechnol.* **19**, 423–428 (2001).
2. Zhao, H., Chockalingam, K. & Chen, Z. Directed evolution of enzymes and pathways for industrial biocatalysis. *Curr. Opin. Biotechnol.* **13**, 104–110 (2002).
3. Joyce, G. F. Directed evolution of nucleic acid enzymes. *Annu. Rev. Biochem.* **73**, 791–836 (2004).
4. Wilson, D. S. & Szostak, J. W. *In vitro* selection of functional nucleic acids. *Annu. Rev. Biochem.* **68**, 611–647 (1999).
5. McPherson, M. J. & Møller, S. G. *PCR* (Springer, New York, 2000).
6. Dykxhoorn, D. M., Novina, C. D. & Sharp, P. A. Killing the messenger: short RNAs that silence gene expression. *Nature Rev. Mol. Cell Biol.* **23**, 1961–1967 (2003).
7. Novina, C. D. & Sharp, P. A. The RNAi revolution. *Nature* **430**, 161–164 (2004).
8. Khudiyakov, Y. E. & Fields, H. A. *Artificial DNA: methods and applications* (CRC, Boca Raton, Florida, 2002).
9. Muller, S., Wolf, J. & Ivanov, S. A. Current strategies for the synthesis of RNA. *Curr. Org. Syn.* **1**, 293–307 (2004).
10. Milligan, J. F. & Uhlenbeck, O. C. Synthesis of small RNAs using T7 RNA polymerase. *Methods Enzymol.* **180**, 51–62 (1989).
11. Watson, J. D. & Crick, F. H. C. Molecular structure of nucleic acids. *Nature* **171**, 737–738 (1953).
12. Uhlenbeck, O. C. A small catalytic oligoribonucleotide. *Nature* **328**, 596–600 (1989).
13. Nissen, P., Hansen, J., Ban, N., Moore, P. B. & Steitz, T. A. The structural basis of ribosome activity in peptide bond synthesis. *Science* **289**, 920–930 (2000).
14. Steitz, T. A. & Moore, P. B. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. *Trends Biochem. Sci.* **28**, 411–418 (2003).
15. Stein, C. A. & Krieg, A. M. *Applied Antisense Oligonucleotide Technology*. (eds Stein, C. A. & Krieg, A. M.) (Wiley, New York, 1998).
16. Crooke, S. T. Progress in antisense technology. *Annu. Rev. Med.* **55**, 61–95 (2004).
17. Reese, C. B. & Yan, H. B. Solution phase synthesis of ISIS 2922 (Vitravene) by the modified H-phosphonate approach. *J. Chem. Soc. Perkins Trans.* **1**, 2619–2633 (2002).
18. Holmlund, J. T. Applying antisense technology. *Ann. NY Acad. Sci.* **1002**, 244–251 (2003).
19. Meister, G. & Tuschl, T. Mechanisms of gene silencing by double-stranded RNA. *Nature* **431**, 343–349 (2004).
20. Eaton, B. E. & Pieken, W. A. Ribonucleosides and RNA. *Annu. Rev. Biochem.* **64**, 837–863 (1995).
21. Sarafian, S. G., Hughes, S. H. & Arnold, E. Designing anti-AIDS drugs targeting the major mechanism of HIV-1 RT resistance to nucleoside analog drugs. *Int. J. Biochem. Cell Biol.* **36**, 1706–1715 (2004).
22. Gold, L., Polisky, B., Uhlenbeck, O. & Yarus, M. Diversity of oligonucleotide functions. *Annu. Rev. Biochem.* **64**, 763–797 (1995).
23. Osborne, S. E. & Ellington, A. D. Nucleic acid selection and the challenge of combinatorial chemistry. *Chem. Rev.* **97**, 349–370 (1997).
24. Koizumi, M., Soukup, G. A., Kerr, J. N. Q. & Breaker, R. R. Allosteric selection of ribozymes that respond to the second messengers cGMP and cAMP. *Nature Struct. Biol.* **6**, 1062–1071 (1999).
25. Soukup, G. A., DeRose, E. C., Koizumi, M. & Breaker, R. R. Generating new ligand-binding RNAs by affinity maturation and disintegration of allosteric ribozymes. *RNA* **7**, 524–536 (2001).
26. Jayasena, S. D. Aptamers: an emerging class of molecules that rival antibodies in diagnostics. *Clin. Chem.* **9**, 1628–1650 (1999).
27. Brockstedt, U., Uzarowska, A., Montpetit, A., Pfau, W. & Labuda, D. *In vitro* evolution of RNA aptamers recognizing carcinogenic aromatic amines. *Biochem. Biophys. Res. Commun.* **313**, 1004–1008 (2004).
28. Sayer, N. M. *et al.* Structural determinants of conformationally selective, prion-binding aptamers. *J. Biol. Chem.* **279**, 13102–13109 (2004).
29. Romig, T. S., Bell, C. & Drolet, D. W. Aptamer affinity chromatography: combinatorial chemistry applied to protein purification. *J. Chromatogr. B. Biomed. Sci. Appl.* **731**, 275–284 (1999).
30. Deng, Q., German, I., Buchanan, D. & Kennedy, R. T. Retention and separation of adenosine and analogues by affinity chromatography with an aptamer stationary phase. *Anal. Chem.* **73**, 5415–5421 (2001).
31. Hermann, T. & Patel, D. J. Adaptive recognition by nucleic acid aptamers. *Science* **287**, 820–825 (2000).
32. Hamaguchi, N., Ellington, A. & Stanton, M. Aptamer beacons for the direct detection of proteins. *Anal. Biochem.* **294**, 126–131 (2001).
33. McCauley, T. G., Hamaguchi, N. & Stanton, M. Aptamer-based biosensor arrays for detection and quantification of biological macromolecules. *Anal. Biochem.* **319**, 244–250 (2003).
34. Jhaveri, S., Rajendran, M. & Ellington, A. D. *In vitro* selection of signaling aptamers. *Nature Biotechnol.* **18**, 1293–1297 (2000).
35. Peracchi, A. Prospects for antiviral ribozymes and deoxyribozymes. *Rev. Med. Virol.* **14**, 47–64 (2004).
36. Opalinska, J. B. & Gewirtz, A. M. Nucleic acid therapeutics: basic principles and recent applications. *Nature Rev. Drug Disc.* **1**, 503–514 (2002).
37. Lin, Y., Qiu, Q., Gill, C. & Jayasena, S. D. Modified RNA sequence pools for *in vitro* selection. *Nucleic Acids Res.* **22**, 5229–5234 (1994).
38. Beaudry, A., DeFoe, J., Zinnen, S., Burgin, A. & Beigelman, L. *In vitro* selection of a novel nuclease-resistant RNA phosphodiesterase. *Chem. Biol.* **7**, 323–334 (2000).
39. Famulok, M. & Verma, S. *In vivo*-applied functional RNAs as tools in proteomics and genomics research. *Trends Biotechnol.* **20**, 462–466 (2002).
40. Toulmé, J.-J., Di Primo, C. & Boucard, D. Regulating eukaryotic gene expression with aptamers. *FEBS Lett.* **567**, 55–62 (2004).
41. Homann, M. & Göringer, H. U. Uptake and intracellular transport of RNA aptamers in African trypanosomes suggests therapeutic 'piggy-back' approach. *Bioorg. Med. Chem.* **9**, 2571–2580 (2001).
42. Vater, A. & Klusmann, S. Towards third-generation aptamers: spiegelmers and their therapeutic prospects. *Curr. Opin. Drug Disc. Devel.* **6**, 253–261 (2003).
43. Eulberg, D. & Klusmann, S. Spiegelmers: biostable aptamers. *ChemBiochem.* **4**, 979–983 (2003).
44. Nolte, A., Klusmann, S., Bald, R., Erdmann, V. A. & Furste, J. P. Mirror-design of L-oligonucleotide ligands binding to L-arginine. *Nature Biotechnol.* **14**, 1112–1115 (1996).
45. Cox, J. C. & Ellington, A. D. Automated selection of anti-protein aptamers. *Bioorg. Med. Chem.* **9**, 2525–2531 (2001).
46. Sooter, L. J. *et al.* Towards automated nucleic acid enzyme selection. *Biol. Chem.* **9**, 1327–1334 (2001).
47. Cox, J. C. *et al.* Automated acquisition of aptamer sequences. *Comb. Chem. High Throughput Screen.* **4**, 289–299 (2002).
48. Cox, J. C. *et al.* Automated selection of aptamers against protein targets translated *in vitro*: from gene to aptamer. *Nucleic Acids Res.* **30**, e108 (2002).
49. Eulberg, D. & Klusmann, S. Spiegelmer growth factor therapy for neovascular age-related macular degeneration: promises and pitfalls. *Ophthalmology* **110**, 879–881 (2003).
50. Boncler, M. A., Koziolkiewicz, M. & Watala, C. Aptamer inhibits degradation of platelet proteolytically activatable receptor, PAR-1, by thrombin. *Thromb. Res.* **104**, 215–222 (2002).
51. Rusconi, C. P. *et al.* RNA aptamers as reversible antagonists of coagulation factor IXa. *Nature* **419**, 90–94 (2002).
52. Nissen, P., Hansen, J., Ban, N., Moore, P. B. & Steitz, T. A. The structural basis of ribosomal activity in peptide bond synthesis. *Science* **289**, 920–930 (2000).
53. Hansen, J. L. *et al.* The structures of four macrolide antibiotics bound to the large ribosomal subunit. *Mol. Cell* **10**, 117–128 (2002).
54. Hansen, J. L., Moore, P. B. & Steitz, T. A. Structures of five antibiotics bound at the peptidyl transferase center of the large ribosomal subunit. *J. Mol. Biol.* **330**, 1061–1075 (2003).
55. Schlünzen, F. *et al.* Structural basis for the interaction of antibiotics with the peptidyl transferase centre in eubacteria. *Nature* **413**, 814–821 (2001).
56. Bagheri, S. & Kashani-Sabet, M. Ribozymes in the age of molecular therapeutics. *Curr. Mol. Med.* **4**, 489–506 (2004).
57. Kawa, D., Wang, J., Yuan, Y. & Liu, F. Inhibition of viral gene expression by human ribonuclease P. *RNA* **4**, 1397–1406 (1998).
58. Plehn-Dujowich, D. & Altman, S. Effective inhibition of influenza production in cultured cells by external guide sequences and ribonuclease P. *Proc. Natl Acad. Sci. USA* **95**, 7327–7331 (1998).
59. Rangarajan, S., Raj, M. L. S., Hernandez, J. M., Grotewold, E. & Gopalan, V. RNase P as a tool for disruption of gene expression in maize cells. *Biomed. J.* **380**, 611–616 (2004).
60. Byun, J., Lan, N., Long, M. & Sullenger, B. A. Efficient and specific repair of sickle beta-globin RNA by trans-splicing ribozymes. *RNA* **9**, 1254–1263 (2003).
61. Sullenger, B. A. & Gilboa, E. Emerging clinical applications of RNA. *Nature* **418**, 252–258 (2002).
62. Perutka, J., Wang, W. J., Goerlitz, D. & Lambowitz, A. M. Use of computer-designed group II introns to disrupt *Escherichia coli* DExH/D-box protein and DNA helicase genes. *J. Mol. Biol.* **336**, 421–439 (2004).
63. Jarvis, T. C. *et al.* Ribozymes as tools for therapeutic target validation in arthritis. *J. Immunol.* **165**, 493–498 (2000).
64. Wilson, C. & Szostak, J. W. *In vitro* evolution of a self-alkylating ribozyme. *Nature* **374**, 777–782 (1995).
65. Unrau, P. J. & Bartel, D. P. RNA-catalysed nucleotide synthesis. *Nature* **395**, 260–263 (1998).
66. Baskerville, S. & Bartel, D. P. A ribozyme that ligates RNA to protein. *Proc. Natl Acad. Sci. USA* **99**, 9154–9159 (2002).
67. Tang, J. & Breaker, R. R. Structural diversity of self-cleaving ribozymes. *Proc. Natl Acad. Sci. USA* **97**, 5784–5789 (2000).
68. Lazarev, D., Puskarz, I. & Breaker, R. R. Substrate specificity and reaction kinetics of an X-motif ribozyme. *RNA* **9**, 688–697 (2003).
69. Emilsson, G. M. & Breaker, R. R. Deoxyribozymes: new activities and new applications. *Cell. Mol. Life Sci.* **59**, 596–607 (2002).
70. Santoro, S. W. & Joyce, G. F. A general purpose RNA-cleaving DNA enzyme. *Proc. Natl Acad. Sci. USA* **94**, 4262–4266 (1997).
71. Santoro, S. W. & Joyce, G. F. Mechanism and utility of an RNA-cleaving DNA enzyme. *Biochemistry* **37**, 13330–13342 (1998).
72. Santiago, F. S. *et al.* New DNA enzyme targeting Erg-1 mRNA inhibits vascular smooth muscle proliferation and regrowth after injury. *Nature Med.* **5**, 1264–1269 (1999).
73. Santiago, F. S. & Khachigian, L. M. Nucleic acid based strategies as potential therapeutic tools: mechanistic considerations and implications to restenosis. *J. Mol. Med.* **79**, 695–706 (2001).
74. Li, Y. & Breaker, R. R. Phosphorylating DNA with DNA. *Proc. Natl Acad. Sci. USA* **96**, 2746–2751 (1999).
75. Wang, W., Billen, L. P. & Li, Y. Sequence diversity, metal specificity, and catalytic proficiency of metal-dependent phosphorylating DNA enzymes. *Chem. Biol.* **9**, 507–517 (2002).
76. Sreedhara, A., Li, Y. F. & Breaker, R. R. Ligating DNA with DNA. *J. Am. Chem. Soc.* **126**, 3454–3460 (2004).
77. Uhlenbeck, O. C. Keeping RNA happy. *RNA* **1**, 4–6 (1995).
78. Jiang, F., Kumar, R. A., Jones, R. A. & Patel, D. J. Structural basis of RNA folding and recognition in an AMP-RNA aptamer complex. *Nature* **382**, 183–186 (1996).
79. Soukup, G. A. & Breaker, R. R. Relationship between internucleotide linkage geometry and the stability of RNA. *RNA* **5**, 1308–1325 (1999).
80. Tang, J. & Breaker, R. R. Rational design of allosteric ribozymes. *Chem. Biol.* **4**, 453–459 (1997).
81. Breaker, R. R. Engineered allosteric ribozymes as biosensor components. *Curr. Opin. Biotechnol.* **13**, 31–39 (2002).
82. Silverman, S. K. Rube Goldberg goes (ribo)nuclear? Molecular switches and sensors made from RNA. *RNA* **9**, 377–383 (2003).
83. Seetharaman, S., Zivarts, M., Sudarsan, N. & Breaker, R. R. Immobilized switches for the analysis of complex chemical and biological mixtures. *Nature Biotechnol.* **19**, 336–341 (2001).
84. Hesselberth, J. R., Robertson, M. P., Knudsen, S. M. & Ellington, A. D. Simultaneous detection of diverse analytes with an aptazyme ligase array. *Anal. Biochem.* **312**, 106–112 (2003).
85. Vaish, N. K. *et al.* Zeptomole detection of a viral nucleic acid using a target-activated ribozymes. *RNA* **9**, 1058–1072 (2003).
86. Kossen, K. *et al.* High-throughput ribozyme-based assays for detection of viral nucleic acids. *Chem. Biol.* **11**, 807–815 (2004).

87. Srinivasan, J. *et al.* ADP-specific sensors enable universal assay of protein kinase assay. *Chem. Biol.* **11**, 499–508 (2004).
88. Ferguson, A. *et al.* A novel strategy for selection of allosteric ribozymes yields RiboReporter™ sensors for caffeine and aspartame. *Nucleic Acids Res.* **32**, 1756–1766 (2004).
89. Najafi-Shoushtari, S. H., Mayer, G. & Famulok, M. Sensing complex regulatory networks by conformationally controlled hairpin ribozymes. *Nucleic Acids Res.* **32**, 3212–3219 (2004).
90. Hartig, J. S. *et al.* Protein-dependent ribozymes report molecular interactions in real time. *Nature Biotechnol.* **20**, 717–722 (2002).
91. Mandal, M. & Breaker, R. R. Gene regulation by riboswitches. *Nature Rev. Mol. Cell Biol.* **5**, 451–463 (2004).
92. Barrick, J. E. *et al.* New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc. Natl Acad. Sci. USA* **101**, 6421–6426 (2004).
93. Nahvi, A. *et al.* Genetic control by a metabolite binding mRNA. *Chem. Biol.* **9**, 1043–1049 (2002).
94. Winkler, W., Nahvi, A. & Breaker, R. R. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* **419**, 952–956 (2002).
95. Mandal, M., Boese, B., Barrick, J. E., Winkler, W. C. & Breaker, R. R. Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. *Cell* **113**, 577–586 (2003).
96. Mandal, M. & Breaker, R. R. Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nature Struct. Mol. Biol.* **11**, 29–35 (2004).
97. Johansen, L. E., Nygaard, P., Lassen, C., Agerso, Y. & Saxild, H. H. Definition of a second *Bacillus subtilis* *pur* regulon comprising the *pur* and *xpt-pbuX* operons plus *pbuG*, *nupG* (*yxjA*) and *pbuE* (*ydjL*). *J. Bacteriol.* **185**, 5200–5209 (2003).
98. Werstuck, G. & Green, M. R. Controlling gene expression in living cells through small molecule-RNA interactions. *Science* **282**, 296–298 (1998).
99. Grate, D. & Wilson, C. Inducible regulation of the *S. cerevisiae* cell cycle mediated by an RNA aptamer–ligand complex. *Bioorg. Med. Chem.* **9**, 2565–2570 (2001).
100. Harvey, I., Garneau, P. & Pelletier, J. Inhibition of translation by RNA–small molecule interactions. *RNA* **8**, 452–463 (2002).
101. Suess, B. *et al.* Conditional gene expression by controlling translation with tetracycline-binding aptamers. *Nucleic Acids Res.* **31**, 1853–1858 (2003).
102. Hanson, S., Berthelot, K., Fink, B., McCarthy, J. E. G. & Suess, B. Tetracycline-aptamer-mediated translational regulation in yeast. *Mol. Microbiol.* **49**, 1627–1637 (2003).
103. Suess, B., Fink, B., Berens, C., Stenz, R. & Hillen, W. A theophylline responsive riboswitch based on helix slipping controls gene expression *in vivo*. *Nucleic Acids Res.* **32**, 1610–1614 (2004).
104. Sudarsan, N., Barrick, J. E. & Breaker, R. R. Metabolite-binding RNA domains are present in the genes of eukaryotes. *RNA* **9**, 644–647 (2003).
105. Sudarsan, N., Wickiser, J. K., Nakamura, S., Ebert, M. S. & Breaker, R. R. An mRNA structure in bacteria that controls gene expression by binding lysine. *Genes Dev.* **17**, 2688–2697 (2003).
106. Winkler, W. C., Nahvi, A., Roth, A., Collins, J. A. & Breaker, R. R. Control of gene expression by a natural metabolite-responsive ribozyme. *Nature* **428**, 281–286 (2004).
107. Kawasaki, H. & Taira, K. Identification of genes by hybrid ribozymes that couple cleavage activity with the unwinding activity of an endogenous RNA helicase. *EMBO Rep.* **3**, 443–450 (2002).
108. Rhoades, K. & Wong-Staal, F. Inverse Genomics™ as a powerful tool to identify novel targets for the treatment of neurodegenerative diseases. *Mech. Age. Dev.* **124**, 125–132 (2003).
109. Gruenert, D. C. *et al.* Sequence-specific modification of genomic DNA by small DNA fragments. *J. Clin. Invest.* **112**, 637–641 (2003).
110. Walther, W. & Stein, U. Viral vectors for gene transfer: a review of their use in the treatment of human diseases. *Drugs* **60**, 249–271 (2000).
111. Kuan, J. Y. & Glazer, P. M. Targeted gene modification using triplex-forming oligonucleotides. *Methods Mol. Biol.* **262**, 173–194 (2004).
112. Long, M. B., Jones, J. P., Sullenger, B. A. & Byun, J. Ribozyme-mediated revision of RNA and DNA. *J. Clin. Invest.* **112**, 312–318 (2003).
113. Garcia-Blanco, M. A., Baraniak, A. P. & Lasda, E. L. Alternative splicing in disease and therapy. *Nature Biotechnol.* **22**, 535–546 (2004).
114. Gusarov, I. & Nudler, E. The mechanism of intrinsic transcription termination. *Mol. Cell* **3**, 495–504 (1999).
115. Yarnell, W. S. & Roberts, J. W. Mechanism of intrinsic transcription termination and antitermination. *Science* **284**, 611–615 (1999).

Acknowledgements Nucleic acids research in the Breaker laboratory is supported by the David and Lucile Packard Foundation, NIH and NSF.

Competing interests statement The author declares competing financial interests: details accompany the paper on www.nature.com/nature.

Exploring biology with small organic molecules

Brent R. Stockwell

Department of Biological Sciences and Department of Chemistry, Columbia University, 614 Fairchild Center, MC 2406, New York, New York 10027, USA (e-mail: stockwell@biology.columbia.edu)

Small organic molecules have proven to be invaluable tools for investigating biological systems, but there is still much to learn from their use. To discover and to use more effectively new chemical tools to understand biology, strategies are needed that allow us to systematically explore 'biological-activity space'. Such strategies involve analysing both protein binding of, and phenotypic responses to, small organic molecules. The mapping of biological-activity space using small molecules is akin to mapping the stars — uncharted territory is explored using a system of coordinates that describes where each new feature lies.

To understand a system, you need to perturb it. This principle underlies most of the experimental sciences and explains why our depth of understanding of biological systems has been largely determined by the availability of tools that can be used to disrupt them. The development of molecular genetics in the twentieth century advanced our understanding of the molecules that control living systems. Now, molecular genetics allows investigators to eliminate specific proteins by 'knocking out' genes; to increase the concentrations of particular proteins by increasing the number of copies of the corresponding genes or by using a more active promoter on such genes; or to alter the function of a protein by introducing specific mutations in the corresponding gene^{1,2}.

Although these methods have proved to be powerful in model organisms such as *Saccharomyces cerevisiae* and *Drosophila melanogaster*, mammals are more difficult to study using genetic-screening approaches because of their slower rates of reproduction, large physical sizes and large genomes. An alternative approach that has been gaining momentum in recent years is the use of small organic molecules instead of mutations. This approach is referred to as chemical genetics and is used to illuminate the molecular mechanisms underlying biological processes³⁻⁷. Because small molecules can alter the functions of proteins by binding to them and inhibiting or activating their normal functions, they can be used to perturb living systems and to reveal the molecular 'wiring diagrams' of these systems. There have been notable successes using this approach, although technical hurdles remain^{3,4}.

The use of small molecules can complement gene-based methods of perturbing protein function, and in some cases, can offer advantages over such methods. For example, a protein may have several functions in a cell. In the case of a deletion mutation, all these functions are lost. However, it is possible to find small molecules that perturb only one of several functions of a protein, resulting in a level of understanding of protein function that would not be possible through gene-based perturbation⁸. In addition, it is easier to exert temporal control of protein function with small molecules because they can be added to induce an effect and later washed away to return a cell to its wild-type state. Finally, although most small molecules are not drugs, the occasional development of a small molecule into a drug can motivate researchers to use small-molecule tools to study biology.

To fully exploit the potential of chemical genetics, it will be necessary to create collections of small molecules that are suited to modulating the functions of many different proteins. However, each protein class generally requires a different

Box 1

Screening for new ligands

When no ligand for a particular protein is known, screening of chemical libraries is often undertaken in the hope of identifying compounds that bind to the protein with reasonable affinity. Two distinct but complementary approaches can be applied: experimental (usually high-throughput; see Box 2) screening and structure-based virtual screening.

In one type of experimental screening, the protein is expressed and purified and used in a high-throughput screen to find small molecules that bind to it. This can be a time-consuming and expensive endeavour, and for many proteins it can fail to yield an effective ligand. Alternatively, in structure-based virtual screening, an atomic resolution structure of the protein is obtained using X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy. This protein structure is then used in a computer-based experiment to find small molecules predicted to bind to the protein. Using programs such as AutoDock, DOCK, FlexX, FRED, GOLD and Glide, millions of compounds can be examined *in silico* for their propensity to interact with the target protein, and the relative fit of each candidate scored⁶⁶⁻⁷⁰. This virtual screening approach has been used to generate ligands for casein kinase II using DOCK and SCORE⁷¹, and for the BCR-ABL oncoprotein using DOCK⁷². Although this is a useful emerging technology, current success rates are low because it is difficult to predict how small molecules will interact with a protein; there is flexibility in the torsion angles in both the protein and small molecule, causing uncertainty regarding the three-dimensional structure of both. Improvements in the predictive accuracy of such programs will affect virtual screening, and so the discovery of novel protein ligands.

Although these two approaches to ligand discovery are distinct, they can be used together to enhance the chances of finding an active compound. In particular, within the pharmaceutical industry, the use of virtual screening as a 'filter' to select compounds from very large virtual libraries for experimental screening has become increasingly common. This filtering process can use various types of information (for example, the crystal structures of the protein itself), with the aim of enriching the library that is experimentally screened with 'active' structures. Furthermore, computational filters can also be used to remove compounds that have inappropriate properties from the screening library, as discussed in Box 3. A review of this topic is given in ref. 73.

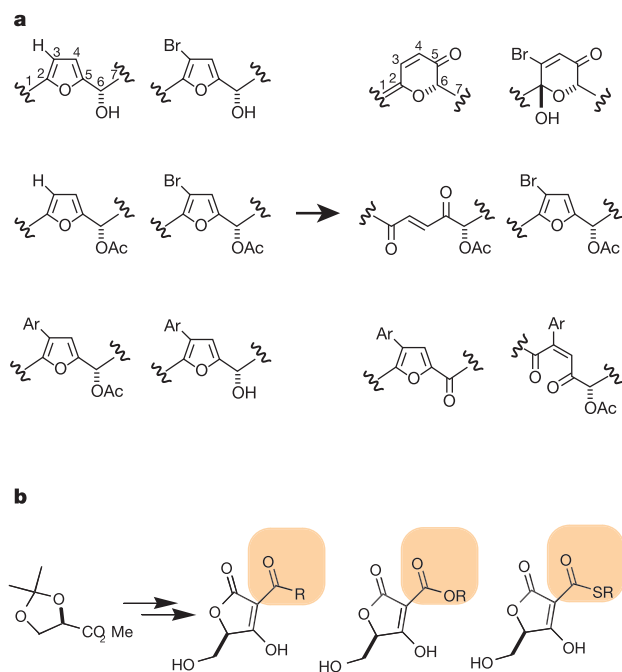


Figure 1 Comparison of diversity-oriented synthesis (DOS) and focused library synthesis (FLS). **a**, The goal of DOS is to create collections of compounds that are maximally diverse, thereby increasing the probability that different proteins will be targeted by different compounds in the library. In the example shown, Burke *et al.*⁶⁵ created a library of compounds with different core structures (skeletons) starting from a common set of precursors (left). The six compounds on the right have different connectivity and are likely to interact with different proteins. **b**, The goal of FLS is to create analogues of the same core structure to optimize binding to a target or class of targets. If the compounds created are too diverse, they may lose their propensity to interact with the designated target protein. In this example, Sodeoka *et al.*¹⁵ created a collection of acyltetronic acids that act as phosphate mimetics and so are likely to inhibit phosphatases. Their synthesis resulted in a library of compounds that are identical except for the portion highlighted in orange.

type of small-molecule modulator. Thus, key aims should be to determine the full range of protein classes that occur in biology and to understand what type of small molecule interacts with each class. A similar argument can be made for determining the full range of phenotypes or observable properties of cells and organisms that occur in biology, given that the molecular basis of phenotypes is what we are ultimately hoping to understand. A central challenge facing the field of chemical genetics is therefore the mapping of ‘biological-activity space’, which involves analysing both protein binding of, and phenotypic responses to, small molecules. My aim here is to describe the challenges — including the design of synthetic chemicals, protein-binding and phenotypic assays, and ensuring quality control — that must be overcome to create a comprehensive map of biological-activity space using small molecules. Other systematic approaches to investigating biological systems, such as the use of RNA interference (RNAi), in which synthetic RNA fragments are designed to interfere with the expression of specific genes⁹, or antibodies¹⁰, are not covered here, but in many cases, could offer complementary information on systems of interest.

Assembling the ‘ideal’ chemical library

If small molecules are to be used as analogues of genetic mutations for studying mammalian systems, they must show the same generality as mutations. That is, they need to be applicable to the study of most or all proteins in an organism⁷. However, the specific chemical structure needed to bind to each protein is necessarily different: the requisite structure is determined by the shape of the available binding pockets

on each protein. So, if we wish to create an ‘ideal’ chemical library for chemical genetics — one that contains a small-molecule ligand or binding partner for each protein — structures that bind to each protein need to be identified.

Of course, no existing chemical library contains compounds that bind selectively to every protein. Furthermore, there are many proteins for which no small-molecule ligand has yet been identified. Identifying new compounds with differing selectivities, or that bind to novel proteins, typically involves some type of screening experiment in which a library of compounds is assessed for the property of interest. Here, I focus primarily on understanding the biological effects of ‘active’ small molecules; that is, those molecules that possess a property of interest. A discussion of the screening approaches used to identify such molecules from the many that have no activity of interest is described in Box 1. The differences between high-throughput screening for modulators of a particular protein (a core activity of the pharmaceutical industry) and performing global analyses of the biological effects of a library of small molecules (a core activity in chemical genetics) are discussed in Box 2.

Whether the goal is to find a ligand for a particular protein, or to use the global effects of a library to elucidate biological processes, the composition of the library used in the screening experiment is a key factor. Libraries can be assembled from available compounds or synthesized *de novo*. In practice, there are two types of chemical libraries that can be synthesized today: ‘focused libraries’ and ‘diversity-oriented libraries’^{11,12} (Fig. 1). Focused libraries are designed around a specific piece of a small molecule, known as a scaffold, and are used to target a specific class of proteins. Often, such scaffolds may be chemically related to endogenous ligands for particular protein classes. Recent examples of focused libraries include those targeted against G-protein coupled receptors (GPCRs)¹³, proteases¹⁴, phosphatases¹⁵ and kinases¹⁶. In contrast, diversity-oriented libraries are

Box 2

High-throughput screens versus global analyses

In a high-throughput screen, many different chemicals (or other test reagents) are evaluated in the same biological test for their effects on a protein or cellular process. The term ‘screen’ is used to indicate that many different chemicals are tested but only a small number of them are expected to be active. The term ‘high-throughput’ is used to indicate that many chemicals are put through this process in a short period of time. There are, however, two types of analyses that can be performed on large data sets: screens and global analyses. Both approaches involve collecting a large amount of data on the effects of specific compounds or other reagents in the same assay. However, the goals of the two approaches differ: screens seek simply to identify several active reagents that can be investigated further in subsequent experiments, whereas global analyses seek to draw meaningful conclusions regarding all the reagents that were tested in the screen. Thus, a high rate of false negatives and false positives can be tolerated in a screen because as long as a few true positives can ultimately be confirmed, the screen is successful. Unfortunately, the same is not true for global analyses, which require low false-positive and false-negative rates for the data to be meaningfully interpreted. Those setting out to perform a global analysis would be wise to consider performing numerous replicates of each compound at several concentrations. It is perhaps only a slight exaggeration to state that academic scientists frequently wish to understand fundamental property relationships between structure and activity, whereas industry scientists often seek to identify a few lead compounds that can be pursued as drug candidates. Nonetheless, the goals of individual researchers performing large numbers of chemical tests differ, and therefore the required data quality, the necessary number of tests and the properties of the tested compounds will be different for each investigator.

Box 3

Additional factors to consider in library design

A number of other properties of small molecules are important to their use as a tool or potential drug, in addition to their ability to bind potently and specifically to particular protein targets. Such properties include their ability to cross biological membranes, to be substrates for drug efflux pumps in cells, their chemical stability, and their solubility in water and dimethyl sulphoxide (a common organic solvent). There has been much interest in the pharmaceutical industry in engineering such 'drug-like' properties and discarding candidate compounds that are unlikely to be effective drugs, even before they are synthesized. The most widely used of these drug-like property rules are those formulated by Lipinski *et al.*, who compared the computed properties of marketed drugs with those of non-drugs⁷⁴.

In recent years, there has been a trend towards creating libraries of compounds that are predicted to be 'lead-like' rather than drug-like. This is in recognition of the fact that as a compound progresses from being a drug lead to an actual drug, its properties tend to change in a consistent way: drugs are typically larger and more hydrophobic than leads⁷⁵. This reflects the practical fact that medicinal chemists tend to add chemical matter rather than remove it during lead optimization. Better predictions of drug-like and lead-like properties will have an important impact on the creation of both drug candidates and chemical tools; chemical tools also need to be soluble, stable and able to penetrate across biological membranes.

not targeted to any specific protein class and are often used in broad screens in which the target proteins are not known. Because the goal of diversity-oriented synthesis (DOS) is to create a maximally diverse collection of compounds, the synthetic planning algorithms required are distinct from those used to create single compounds or focused libraries^{17,18}. Recent examples of DOS include the synthesis of tricyclic compounds using Ferrier and Pauson–Khand reactions with a glycal template¹⁹, and the synthesis of tetrahydroquinoline²⁰ and hydroxyindole²¹ derivatives.

Each approach to chemical-library design has its advantages and disadvantages. Compounds in focused libraries are more likely than random compounds to be active, but they only target proteins in a known class. Diversity-oriented libraries, in contrast, offer the possibility of targeting entirely new classes of proteins, but any individual compound has a lower probability of activity. The pharmaceutical industry, being justifiably risk-averse, has moved towards the use of focused libraries. Practitioners argue that fewer compounds of greater quality and with a greater probability of becoming drugs are more valuable than larger libraries with compounds that are not likely to become drugs (see Box 3 for a discussion of additional factors considered by the pharmaceutical industry when assembling screening libraries, some of which could also be important for libraries for chemical genetics). Some academic groups, however, without the same constraints of industry, are pursuing higher-risk strategies centred on diversity-oriented approaches. The two approaches are ultimately complementary: a ligand to a new protein class discovered from a diversity-oriented library can serve as the basis for a future focused library that explores the structure–function relationships for compounds targeting this new class of proteins.

More effective chemical libraries for chemical genetics would contain compounds that affect specific proteins and phenotypes but not other closely related proteins and phenotypes. These compounds should also collectively affect a diverse range of proteins and phenotypes. The design of more effective libraries would be aided by assessments of the specificity and diversity of existing libraries, and of each new chemical library as it is designed and synthesized. This would mean that optimal libraries for a given purpose could be rationally assembled from members of other libraries.

Chemical-diversity analysis is routinely carried out today using commercial software packages that catalogue the diversity of structures present in a library (Box 4). But more relevant is the diversity of biological activities shown by a library of compounds. For example, consider a library of ten compounds that have dramatically different structures but that all bind to the protein tubulin: this is a library with significant chemical diversity, but minimal diversity of biological activity. Although there is often a correlation between chemical diversity and the diversity of biological activity, there is not a simple one-to-one correspondence.

To assess the biological-activity diversity of a compound library, it is necessary to evaluate the range of biological activities shown by the library. This involves parameterizing 'biological-activity space', or creating 'metrics' that characterize the activity and specificity of each compound in a library. Protein-binding is a useful metric because many small molecules exert their biological effects by interacting with specific proteins in cells. Phenotypic activity is also useful to measure because ultimately we are interested in understanding how protein binding relates to phenotypic changes.

Indeed, such approaches have been implemented by several groups. Kauvar *et al.*²² reported a protein-affinity map of 'chemical space' and showed that the pattern of protein binding by small molecules can be used to cluster compounds. Greenbaum *et al.*²³ used a similar approach, which they termed affinity fingerprinting. They used this approach to characterize the affinity of a library of peptidic epoxides for numerous proteases and thus to group these proteases by reactivity. Finally, Weinstein *et al.*²⁴ used an analogous approach with a phenotypic assay. By measuring the effects of compounds on the proliferation of a panel of 60 tumour cell lines, Weinstein and colleagues²⁴ discovered that compounds with similar structures or similar mechanisms of action had similar phenotypic profiles (that is, inhibited the growth of a similar set of tumour cell lines). In the remainder of this review, I will consider the status of the methods available for further exploring 'biological-activity space' and consider some of the key challenges inherent in this endeavour.

Protein-binding assays

Methods have been created to measure the ability of small molecules to bind to specific proteins²⁵ (Table 1). In recent years, there has been

Box 4

Calculating chemical diversity

Small organic molecules come in all shapes and sizes. The diversity of a library is a quantitative description of how different these compounds are from each other. Consider library A with ten compounds that all look identical except for the nature of one side-chain, compared to library B with ten compounds that have dramatically different sizes and shapes. Intuitively, most people agree that library A is in some way less diverse than library B. However, to be rigorous it is necessary to specify the attributes that are more or less diverse in these two libraries. For example, if we were to calculate the range of molecular masses in the two libraries and to find that library A has molecular masses that range from 300 to 350 daltons but that library B has molecular masses that range from 200 to 500 daltons, we could say that in terms of molecular mass, library B covers six times the range of molecular masses in library A. Similarly, we could calculate the differences in the ranges of other properties, such as charge, number of atoms, number of rotatable bonds and so on. Such properties, called descriptors, can readily be calculated using commercially available software. These descriptors allow for a quantitative description of chemical diversity. Unfortunately, an additional complication is that diversity of chemical structure does not necessarily imply diversity of biological activity. Finding descriptors for biological activity is necessary to describe the diversity of biological activities for compounds present in a library.

a trend towards testing the specificity of a compound for binding one protein relative to related proteins of the same class (for example, kinases)^{26,27}. Such protein-binding assays can be divided into two types: those that use labelled compounds and those that are label-free (a label is a fluorescent or radioactive group that is added to a test compound). Although labels make protein–ligand interactions easier to observe, they can also be difficult to introduce into a compound, which increases the time and expense associated with measuring protein binding. A brief description of the main assay formats of each type can be found in Table 1, together with references that contain further information on each type of assay.

Although methods are available for measuring the binding of a small molecule to a protein or to a handful of related proteins, few methods systematically measure the binding of small molecules to hundreds or thousands of proteins (Box 2). Such high-throughput protein-binding measurements are required if we are to capture the range of activities shown by small molecules. Label-free detection methods are preferred because they do not require the extra synthetic chemistry involved in introducing a label, and because introducing a label may change the properties of a molecule. However, such measurements can be more difficult to perform: without a label, a larger amount of both protein and compound must often be produced, and the instruments used for label-free measurements are slow (Table 1).

Recent attempts to create high-throughput assays for measuring protein–ligand interactions require the use of labels. One class of high-throughput assay involves immobilizing each test compound on a surface and then incubating these immobilized compounds with a soluble labelled protein^{28–30}. Many compounds can be immobilized side by side on a surface, so this method can measure thousands of protein–small-molecule interactions (Fig. 2a). Kuruvilla *et al.*⁸ used this technique successfully to screen 3,780 compounds for those that bind to the transcriptional repressor Ure2p, and found a compound that disrupts one of the functions of Ure2p.

A related method involves immobilizing compounds on a surface and then detecting the binding of a protein to each compound using

surface plasmon resonance³¹ (Table 1). These surface-based methods can be useful for measuring the ability of many compounds to bind to one or several proteins. For example, Birkert *et al.* used such a method to measure the binding of immobilized triazines to antibodies and to screen 384 compounds for those that act as thrombin inhibitors^{32,33}.

It is possible to invert these surface-based methods and to immobilize thousands of proteins side by side on a surface^{34,35}. A small molecule with a label, such as a fluorescent or radioactive group, can be applied to the surface, washed away and detected by measuring the remaining label (Fig. 2b). Some applications of protein microarrays include the use of an array of most yeast proteins to assess the global pattern of protein activities found in yeast cells³⁶, the discovery of novel protein–protein interactions in human cells³⁷ and an analysis of interactions between human 49 leucine zipper transcription factors³⁸.

A variation of this technology involves creating arrays of expression plasmids, which encode the information required to produce each protein of interest. Creating DNA arrays has become routine in the past decade and is preferable to creating arrays of proteins directly, primarily because DNA can be amplified and because thousands of different DNA-expression constructs will have similar chemical properties (solubility, stability, and so on). In contrast, thousands of different proteins will show idiosyncratic properties that are unique to each protein. It is possible to either place cells on this DNA array and cause proteins to be produced inside the cells³⁹, or to use a cell lysate (produced from cells that have been broken open) to produce an array of proteins *in vitro* (Fig. 2c)⁴⁰. In either case, the net result is a protein array without the added complication of purifying and immobilizing each protein. However, post-translational modifications and protein complexes that are physiologically relevant will not be captured in these formats. So far, only proof-of-principle experiments have been performed with these more recent technologies.

A final high-throughput method for measuring the binding of many proteins to one or more small molecules also has the advantage of not requiring protein purification. This is the three-hybrid system, which is typically carried out in yeast or bacterial cells⁴¹. In such

Table 1 Methods for measuring the affinity of small molecules for proteins

Method	Label	Throughput	Phase	Protein amount	Description	References
Fluorescence polarization	Yes	High	Solution	High	Measures the change in tumbling rate for a compound when it is bound to a protein using loss-of-polarization of incident light.	57
Fluorescence perturbation	No	Low	Solution	High	Measures the change in fluorescence (usually of a tryptophan residue) on a protein caused by proximity of a bound compound.	58
Fluorescence correlation spectroscopy	Yes	High	Solution	Medium	Measures the correlated changes in fluorescence properties that occur when a labelled compound is bound to a labelled protein.	59
Radioligand binding assay	Yes	High	Solid	High	Measures retention of a compound on a surface holding a protein. Detection is by means of a radioactive isotope that is incorporated into the compound.	60
Nuclear magnetic resonance	No	Low	Solution	High	Measures changes in the nuclear spin of protons on a protein that occur when a compound is bound nearby.	61
Mass spectrometry	No	Medium	Solution	Medium	Measures the change in molecular weight observable in a mass spectrometer when a compound is bound to a protein.	62
Surface plasmon resonance	No	Low/medium	Solid	High	Measures the change in refractivity of a metal surface when a compound binds to a protein that is immobilized on that surface.	63
Isothermal titration calorimetry	No	Low	Solution	High	Measures the very small amount of heat produced when a compound binds to a protein.	64
Differential scanning calorimetry	No	Low	Solution	High	An alternative method of measuring the heat released when a compound binds to a protein.	64
Small-molecule microarray	Yes	High	Solid	Medium	Compounds are immobilized on a surface. A protein is incubated with the surface and localizes to where the protein-interacting compounds are bound.	28–30
Protein microarray	Yes	High	Solid	Low/none	Proteins are immobilized on a surface and a labelled compound is incubated with the surface. The compound localizes to where the compound-binding proteins are bound.	34
Three-hybrid system	Yes	High	In cells	None	A compound is covalently linked to an anchor compound that interacts with a DNA-binding protein. If the test compound can bind to a test protein that is fused to a DNA activation domain, transcription of a reporter gene is initiated.	41

systems, a test protein is fused to the activation domain of a transcription activator, and the test small molecule is synthetically linked to an 'anchor' compound that will interact with a protein containing a DNA-binding domain (Fig. 2d). So, if the test small molecule is able to interact with the test protein, the transcription activation domain will be brought into close proximity with the DNA-binding domain, and expression of the reporter gene that is controlled by the system will be activated. This method was used successfully by Liberles *et al.*⁴² to create a mutant version of the FKBP-rapamycin binding domain (FRB), which binds to a modified, non-toxic version of rapamycin.

Although several high-throughput methods have been developed for measuring protein–ligand interactions, many desirable features are not found in these systems. First, measuring the binding of small

molecules to target proteins in solution is preferable to using a surface-based method that may interfere with protein–ligand binding. Unfortunately, most high-throughput methods involve immobilizing either the ligand or the protein on a solid surface to allow parallel processing of all samples with a single solution. Second, it is better to avoid the use of labels on both small molecules and proteins because of the added time and expense needed to introduce such labels into thousands of compounds or proteins, and because the labels may change the activity of the compound or protein. Third, it is easier to use only minute quantities of protein, or better still, to manipulate only the corresponding DNA sequences and allow the system to produce the desired proteins *in situ*. This obviates the need to purify many proteins, each with their own solubility requirements. Fourth, it would be useful to have a system that is 'scalable', both in terms of the binding of small molecules and of the proteins; ideally, it should be possible to automate the detection of the binding of thousands of proteins to thousands of ligands without the need for idiosyncratic modifications to the system for each ligand or each protein. Finally, all these technologies require significant investments in capital equipment and knowledge bases, which limit their adoption by many users. Thus, although each of these problems may ultimately be solved, significant barriers will prevent the widespread adoption of these technologies in the near future.

Phenotypic outcomes

In assessing the biological activities of small molecules, it is useful to consider not only protein binding but also phenotypic effects. Cellular phenotypes that are affected by small molecules include varied phenomena, such as cell death, cell migration, cell proliferation, gene expression, vesicle sorting and axonal sprouting. Organismal phenotypes affected include body weight, tumour formation, joint inflammation and the capacity for learning and memory, among many others. In fact, although there is a finite number of proteins within a given organism, theoretically an infinite number of phenotypes may be assessed for an organism. Given the infinite number of phenotypes that can exist, phenotypic assays performed for library-assessment purposes need to be prioritized in some way. Usually, this prioritization is based on ease of measurement.

It is useful to consider how phenotypic measurements can be automated and undertaken in a high-throughput fashion to characterize the biological activity and specificity of chemical libraries. Most phenotypic measurements cannot be performed in high-throughput simply because they involve time-consuming measurements that use whole organisms, such as mice, worms, flies or zebrafish. In fact, measuring the effect of a single compound on a phenotype in mice typically involves several months of work and costs tens of thousands of dollars. For example, my laboratory recently discovered a compound, indoprofen, which has potential relevance to the pediatric genetic disease spinal muscular atrophy (SMA)⁴³. To test this compound on mouse SMA phenotypes, we needed to evaluate potential routes of administration, achievable concentrations in the plasma, brain and *in utero* embryos given various doses, and toxicity and teratogenicity in pregnant mice. Although we found that this compound had a modest effect at extending the survival of embryos with an SMA genotype, such assays are expensive and time-consuming to perform. Such phenotypic measurements can be valuable for specific compounds of interest but they are not compatible with assessing the activity and specificity of large compound libraries. For this purpose, high-throughput phenotypic assays are needed (Fig. 3).

A number of high-throughput phenotypic assays have been developed, including assays that measure cell viability or proliferation^{3–5}. Such assays measure the presence of intact cell membranes, the abundance of cellular energy (ATP concentration), or the presence of cellular reductases or esterases, which are found in nearly all cells. Such viability assays have been extended to the analysis of synthetic lethal effects: a compound is tested for its ability to kill cells in the

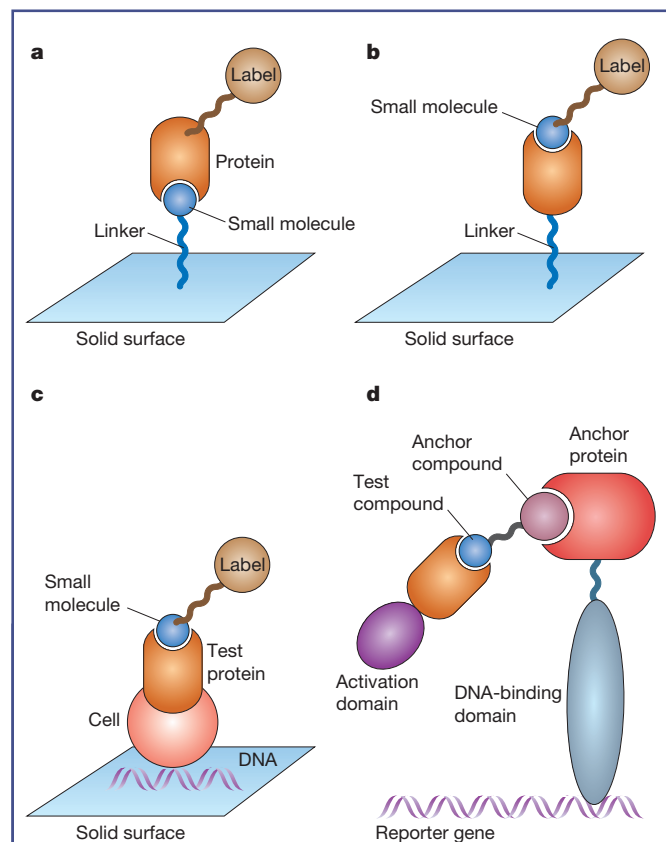


Figure 2 High-throughput-assay formats for detecting small molecule–protein interactions. **a**, Small molecules can be covalently linked to a surface. Meanwhile, a test protein in solution is brought into contact with the surface. The protein binds to small molecules on the surface with high affinity. If the protein is tagged with a label, these interactions can be detected. **b**, Proteins can similarly be immobilized on a surface and brought into contact with a labelled small molecule in solution. High-affinity interactions between the small molecule and specific proteins can then be detected by imaging the locations to which the small molecule binds. **c**, DNA expression plasmids can be arrayed on a surface and cells subsequently plated on top of these expression plasmids. The cells take up the DNA and produce the proteins encoded by each plasmid. Thus, this method allows for the creation of a microarray of cells that overexpress defined proteins. When a labelled compound is brought into close proximity of the array, it localizes to where cells are overexpressing these high-affinity compound-binding proteins. **d**, Yeast three-hybrid system. Transcription factors that regulate gene expression can be divided into DNA-binding domains and transcription-activation domains. It is possible to fuse the complementary DNA sequence of a DNA-binding domain to the cDNA of an anchor protein that interacts with a known small molecule (anchor compound). The anchor compound is then chemically fused to a new test compound. If the cDNA of an activation domain is fused to the cDNA of a test protein, it is possible to determine whether the test protein interacts with the test compound with high affinity by determining whether transcription of a reporter gene has been activated.

presence, but not in the absence, of a defined element, such as another compound or a gene of interest⁴⁴. Identifying compounds that have genotype-selective activity is of interest both because such compounds can be developed into safer drugs with fewer side effects and because they can reveal the molecular consequences of oncogenic mutations in tumour cells. Moreover, viability assays can be used to search for chemical suppressors; a compound is tested for its ability to prevent the lethality of another compound or a toxic gene product. For example, Wang and Dreyfuss⁴⁵ screened for compounds that

prevent the cell death that occurs when the survival motor neuron (SMN)-gene protein is eliminated from mammalian cells. Similarly, Aiken *et al.*⁴⁶ screened for compounds that prevent apoptotic cell death caused by the mutant huntingtin protein in PC12 cells.

Recently, gene-expression signatures have been developed into high-throughput, phenotypic assays⁴⁷. In this approach, a gene-expression profile is measured using DNA microarrays for two cell states of interest, such as undifferentiated neutrophil (a type of granular white blood cell) precursors and differentiated neutrophils.

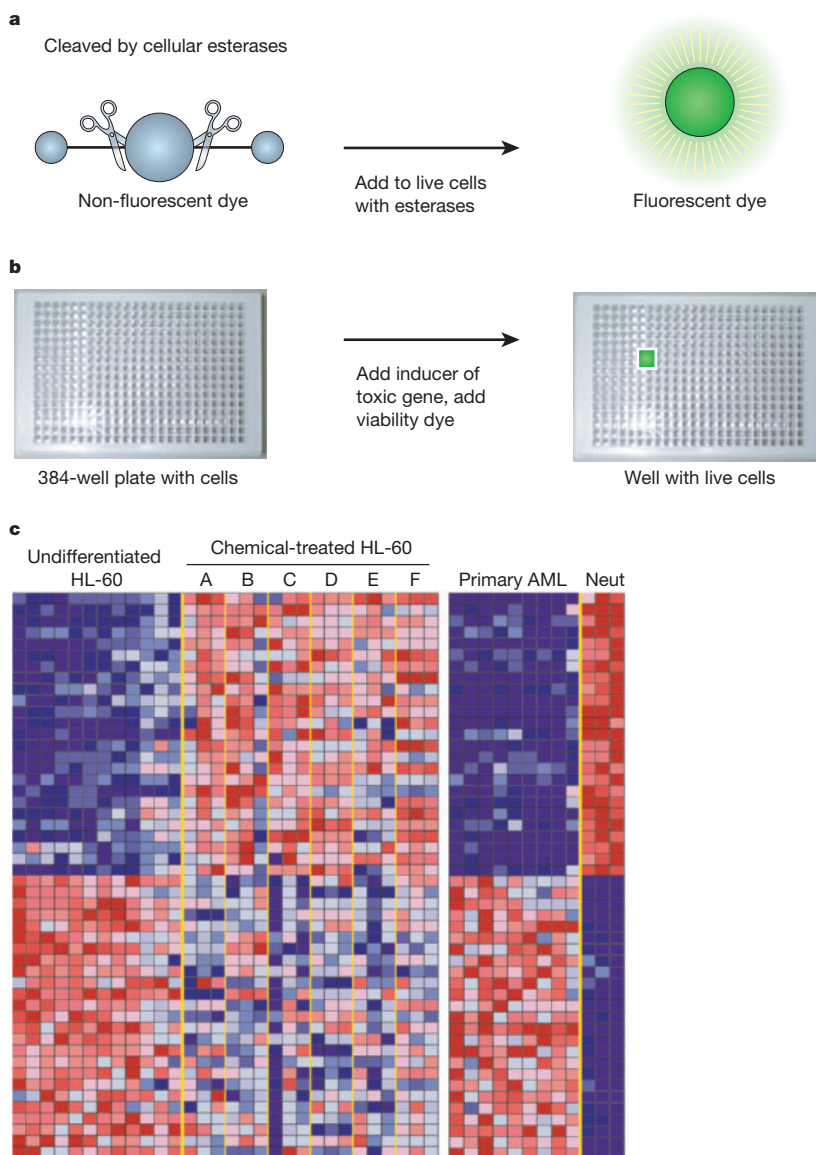


Figure 3 Examples of high-throughput phenotypic screens. These are measurements of properties of cells that can be performed in a parallel fashion and so allow for the testing of many different chemicals at once. **a**, Fluorescence-based viability can be used to measure the number of living cells in a miniaturized test tube. The non-fluorescent dye calcein acetoxymethyl ester, shown schematically in blue, can be cleaved by intracellular esterases to create a fluorescent compound (shown in green). **b**, Such a dye can be used to measure the number of live cells in 384-well plates, which hold 384 individual miniature chambers for growing cells. For example, if a toxic gene is introduced, cells will die unless they are treated with a chemical that is able to prevent this cell death. In this example, the wells holding cells treated with such a chemical are bright green because the viability dye becomes fluorescent on being cleaved by esterases from live cells. **c**, A pattern of gene expression can be used as a signature of the state of a cell. In this example by

Stegmeier *et al.*⁴⁷, gene-expression signatures were obtained for: (1) human neutrophil precursors (HL-60 tumour cells, left) that have failed to differentiate and have become tumour cells; (2) primary acute myelogenous leukaemia (AML) cells from patients (right); and (3) differentiated human neutrophils (Neut, far right). A screen was performed to identify compounds that convert the signature of the HL-60 tumour cell line into the signature of differentiated neutrophils, with the goal of rendering the HL-60 tumour cells non-tumorigenic. Six compounds (of approximately 2,000 tested) were found to induce this switch in gene signatures (labelled 'Chemical-treated HL-60, A to F'). Each row in this table shows the expression level of a different gene under these different conditions (the columns). The colour indicates whether expression in the sample is high (red) or low (blue). The six compounds shown revert the gene-expression pattern of HL-60 tumour cells to that of differentiated neutrophils.

	K1	K2	K3	K4	K5	K6	K7	K8	K9
Compound 1	10e-6	10e-6	10e-2	10e-2	10e-2	10e-2	10e-2	10e-2	10e-2
Compound 2	10e-6	10e-2	10e-2	10e-2	10e-2	10e-2	10e-2	10e-2	10e-2
Compound 3	10e-2	10e-2	10e-8	10e-2	10e-2	10e-2	10e-2	10e-2	10e-2
Compound 4	10e-2	10e-2	10e-2	10e-9	10e-2	10e-2	10e-2	10e-2	10e-2
Compound 5	10e-2	10e-2	10e-2	10e-2	10e-6	10e-2	10e-2	10e-2	10e-2
Compound 6	10e-9	10e-6	10e-8	10e-2	10e-2	10e-2	10e-2	10e-2	10e-2
Compound 7	10e-8	10e-2	10e-2	10e-2	10e-2	10e-2	10e-7	10e-2	10e-2
Compound 8	10e-2	10e-2	10e-2	10e-2	10e-2	10e-8	10e-8	10e-9	10e-6
Compound 9	10e-2	10e-2	10e-2	10e-2	10e-2	10e-2	10e-2	10e-2	10e-6

Figure 4 Using biological-activity matrices to determine the proteins that regulate phenotypes. A hypothetical activity matrix for a library of nine kinase inhibitors. Each row lists the affinity (that is, the equilibrium dissociation constant, written in scientific notation, where 10e-6 represents 0.000001 M) of one compound for each of nine different kinase proteins. Smaller numbers indicate higher affinity. The affinities less than or equal to 10e-6 are highlighted in red because these correspond to high-affinity compounds for these targets. The kinase proteins are labelled K1 to K9. The same affinity matrix can be used to determine which kinases are involved in specific biological processes. In this hypothetical example, if the four compounds highlighted in blue are all capable of inhibiting the growth of a tumour cell line, the K1 kinase is probably responsible for the ability of these compounds to inhibit the growth of this cell line: this is the only kinase to be targeted by all four compounds.

Then the profiles are compared and a gene signature is created which determines whether the cell is in one state or the other. By measuring the effects of small molecules on the appearance of this gene signature, it is possible to determine whether each compound changes the cell state (for example, induces differentiation of neutrophil precursors into neutrophils).

Another emerging trend in high-throughput phenotypic assays involves imaging cells using an automated microscope⁴⁸. Such an approach allows for the detection of phenotypes that can be measured using microscopy. For example, Yarrow *et al.*⁴⁹ recently used an imaging-based screen to identify compounds that affect cell migration during wound healing; Kau *et al.*⁵⁰ used this technique to screen for compounds that prevent nuclear export of FOXO transcription factors. Image-analysis algorithms then allow for the automated processing of these images so that conclusions regarding the effects of compounds on these phenotypes can be extracted. Imaging-based phenotypes could allow for the digitization and clustering of otherwise unrelated phenotypes. Because any image consists of a series of pixels with distinct values, the relationship between any two images can be quantified mathematically.

Finally, the concentration of a particular messenger RNA or protein, such as the SMN protein, can represent a phenotype of interest. For example, patients with the disease SMA have a low SMN protein phenotype. Finding mechanisms and compounds that convert these cells to producing abundant SMN protein is of interest. This concept of molecular phenotypes can be extended to include the measurements of thousands of proteins or mRNAs simultaneously. The global pattern of these proteins or mRNAs represents a quantifiable state of a cell. Thus, measuring the abundance of thousands of proteins, mRNAs or metabolites can be used to create cell signatures or phenotype measurements. Unfortunately, it is not yet feasible to perform such global measurements of protein, mRNA or metabolite abundance in high-throughput. Moreover, some phenotypes do not involve significant transcriptional changes, whereas others do not involve significant changes in protein or metabolite concentrations. New methods for automating and rapidly performing such measurements would be of value.

Creation and use of biological-activity matrices

After collecting a large amount of data on the ability of the members of a chemical library to bind to a set of proteins and affect a set of phenotypes, the data can be analysed to determine the relationship between chemical structure and biological activity. Each compound can be assigned a vector that describes the quantitative level of binding to each protein, and the quantitative effect this has on each phenotype. Comparing these parameters for different libraries could reveal how specific scaffolds and functional elements influence specificity and diversity. Figure 4 shows an example of how a compound might be evaluated for its ability to bind to nine different kinases. Although this evaluation has not been performed, it should be straightforward to do so.

Such data sets can be used to generate hypotheses regarding the molecular mechanisms underlying biological phenotypes⁵¹. For example, if each compound in a library has been annotated with a pattern of protein-binding activity, then it is possible to determine whether binding to any specific protein is correlated with the ability to induce a phenotypic change. In validating such an approach, Root *et al.*⁵¹ rediscovered that small molecules that bind to tubulin are highly likely to inhibit tumour-cell proliferation. This approach can be extended to targets other than proteins: Root *et al.*⁵¹ also found that compounds that bind to small ions, such as potassium, are able to selectively inhibit the proliferation of lung tumour cells relative to other cells. By annotating compound libraries with high-quality target binding and phenotypic profiles, it is possible to extract information regarding the molecules that regulate these phenotypes.

Further challenges

Specificity of small molecules

One limitation of small molecules is their frequent lack of specificity for a single target protein. This can be problematic when using small molecules both as therapeutic agents and as chemical probes: a lack of specificity can lead to unexpected toxicity, preventing the development of an otherwise promising compound into a drug, and can also confound interpretation of the effects of a compound. This problem of non-specificity is often dose-dependent: at higher concentrations, compounds interact with additional proteins. In addition, specific functional groups and scaffolds have been found to be promiscuous, in the sense that they allow binding to a wide range of proteins or non-specific killing of a wide-range of cell types⁵². Such chemical functions need to be identified and removed from future library designs.

There are several strategies for overcoming the problem of specificity. First, it is preferable to identify and use potent compounds (that is, compounds that are likely to modulate a target protein at low nanomolar or picomolar concentrations) because at such low concentrations they are less likely to affect other proteins. Second, measuring the binding specificity of compounds in the type of large-scale protein-binding assays described above should identify some of the alternative protein targets of compounds. Third, it is always critical to confirm the putative mechanism of action of a compound using either additional compounds or other reagents, such as small interfering RNAs (siRNAs)^{53,54}. Although the phenotypic consequences of an RNAi reagent and a small molecule targeting the corresponding mRNA are not always the same, their effects are often sufficiently similar to make this comparison useful. RNAi itself can lack specificity, and it is necessary to test numerous RNAi reagents designed against a target mRNA sequence⁵⁵. Finally, a large collection of RNAi reagents can be a useful tool for high-throughput screens⁹. By using such collections, it should eventually be possible to measure the phenotypic consequences of turning off expression of each gene in an organism.

Building redundancy into a set of probe molecules is an effective way of dealing with the problem of specificity. That is, it is desirable to have not just one compound that inhibits each protein, but rather dozens of compounds that inhibit each protein. If inhibition of

protein X causes phenotype Y, we would expect — in an ideal world — all the small molecules in our collection that inhibit protein X to cause phenotype Y. In the real world, not every protein-X inhibitor will be effective, because some will bind protein X in slightly different ways or be metabolized differently in different cell types. Nonetheless, our confidence that the modulation of protein X causes phenotype Y should be proportional to the percentage of our protein-X inhibitors that cause phenotype Y. Thus, the problem of specificity can be overcome by assembling a sufficiently redundant set of probe compounds: even if no single compound is specific for one target protein, the collection as a whole contains the requisite information on the effects of modulating each target protein.

Finally, given that compounds have different specificities at different concentrations, it would be preferable to collect information on the effects of each compound at multiple concentrations; a full dose-response curve for each compound would be ideal. Unfortunately, the added time and expense associated with collecting this additional information usually makes it impractical. Therefore, new technologies that allow an increase in the number of tests performed per unit time would be valuable. Alternatively, a smaller number of compounds may be tested with more replicates and a full dose-response curve. This trade off between the number of compounds tested and the quality and completeness of the data set collected for each compound needs to be optimized in each project.

Quality control

When collecting large-scale data sets, attention to quality control is crucial. However, there is an inherent trade off between the level of throughput and data quality in large-scale data collection. A minimum level of quality is necessary to ensure that reliable conclusions are extracted from such data sets. However, attention to data quality has not been a priority for many researchers engaged in high-throughput chemical screens, simply because the data quality required for a screen is much lower than the data quality required for a global analysis⁵⁶ (Box 2).

In addition, it is important to eliminate artefacts through the use of counter screens for properties that could interfere with the assay readout, such as intrinsic compound fluorescence or compound aggregation. In general, a counter screen is performed on the compounds that emerge from an initial screen, and compounds that are active in the counter screen are not taken further. For example, in a screen that uses the fluorescent dye calcein as a detection method (Fig. 3), any compound that shows the same colour of fluorescence as calcein will appear to be a positive compound from the screen; a counter screen would involve testing each compound for its intrinsic fluorescence to eliminate those compounds that were falsely active because of this property.

Finally, it is important to assess the solubility and stability of each tested compound or protein, and to confirm that the chemical being tested is the desired one. Solubility can be measured using nephelometry, which detects insoluble particles in solution, and compound identities can be confirmed using liquid chromatography and mass spectrometry. All these methods of improving data quality increase the time and expense associated with large-scale data collection but are crucial if meaningful conclusions are to be drawn.

Outlook

Designing better tools with which to perturb biological systems requires a systematic evaluation of the properties of existing tools. Although large-scale measurements of the effects of small molecules on proteins and phenotypes can be challenging, the resulting data sets can be useful in probing biological-activity diversity. New ways to increase the complexity and sophistication of the phenotypic assays and protein-binding measurements that can be performed on vast arrays of molecules will prove valuable. Moreover, more comprehensive and effective compound libraries will allow us to perturb an increasing percentage of the macromolecules that make

up living systems. In so doing, we may move closer to understanding the roles of the diverse molecules that are responsible for life, death and disease. □

doi:10.1038/nature03196

- Brenner, S. The genetics of *Caenorhabditis elegans*. *Genetics* **77**, 71–94 (1974).
- Hartwell, L. H. Twenty-five years of cell cycle genetics. *Genetics* **4**, 975–80 (1991).
- Stockwell, B. R. Chemical genetics: ligand-based discovery of gene function. *Nature Rev. Genet.* **1**, 116–25 (2000).
- Stockwell, B. R. Frontiers in chemical genetics. *Trends Biotechnol.* **18**, 449–455 (2000).
- Stockwell, B. R. Chemical genetic screening approaches to neurobiology. *Neuron* **36**, 559–562 (2002).
- Schreiber, S. L. The small-molecule approach to biology: chemical genetics and diversity-oriented organic synthesis make possible the systematic exploration of biology. *Chem. Eng. News* **81**, 51–61 (2003).
- Schreiber, S. L. Chemical genetics resulting from a passion for synthetic organic chemistry. *Bioorg. Med. Chem.* **6**, 1127–1152 (1998).
- Kuruvilla, F. G., Shamji, A. F., Sternson, S. M., Hergenrother, P. J. & Schreiber, S. L. Dissecting glucose signalling with diversity-oriented synthesis and small-molecule microarrays. *Nature* **416**, 653–657 (2002).
- Hannon, G. J. RNA interference. *Nature* **418**, 244–251 (2002).
- Moore, P. & Clayton, J. To affinity and beyond. *Nature* **426**, 725–731 (2003).
- Schreiber, S. L. Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science* **287**, 1964–1969 (2000).
- Young, S. S. & Ge, N. Design of diversity and focused combinatorial libraries in drug discovery. *Curr. Opin. Drug Discov. Dev.* **7**, 318–324 (2004).
- Jimonet, P. & Jager, R. Strategies for designing GPCR-focused libraries and screening sets. *Curr. Opin. Drug Discov. Dev.* **7**, 325–333 (2004).
- Reid, R. C. et al. Countering cooperative effects in protease inhibitors using constrained beta-strand-mimicking templates in focused combinatorial libraries. *J. Med. Chem.* **47**, 1641–1651 (2004).
- Sodeoka, M. et al. Synthesis of a tetronic acid library focused on inhibitors of tyrosine and dual-specificity protein phosphatases and its evaluation regarding VHR and cdc25B inhibition. *J. Med. Chem.* **44**, 3216–3222 (2001).
- Stahura, F. L., Xue, L., Godden, J. W. & Bajorath, J. Molecular scaffold-based design and comparison of combinatorial libraries focused on the ATP-binding site of protein kinases. *J. Mol. Graph Model* **17**, 1–9, 51–2 (1999).
- Burke, M. D. & Schreiber, S. L. A planning strategy for diversity-oriented synthesis. *Angew. Chem. Int. Edn Engl.* **43**, 46–58 (2004).
- Spring, D. R. Diversity-oriented synthesis: a challenge for synthetic chemists. *Org. Biomol. Chem.* **1**, 3867–3870 (2003).
- Kubota, H., Lim, J., Depew, K. M. & Schreiber, S. L. Pathway development and pilot library realization in diversity-oriented synthesis: exploring Ferrier and Pauson-Khand reactions on a glycol template. *Chem. Biol.* **9**, 265–276 (2002).
- Couve-Bonnaire, S., Chou, D. T., Gan, Z. & Arya, P. A solid-phase, library synthesis of natural-product-like derivatives from an enantiomerically pure tetrahydroquinoline scaffold. *J. Comb. Chem.* **6**, 73–77 (2004).
- Arya, P., Wei, C. Q., Barnes, M. L. & Daroszewski, M. A solid phase library synthesis of hydroxyindoline-derived tricyclic derivatives by Mitsunobu approach. *J. Comb. Chem.* **6**, 65–72 (2004).
- Kauvar, L. M., Villar, H. O., Sportsman, J. R., Higgins, D. L. & Schmidt, D. E. J. Protein affinity map of chemical space. *J. Chromatog. B* **715**, 93–102 (1998).
- Greenbaum, D. C. et al. Small molecule affinity fingerprinting. A tool for enzyme family subclassification, target identification, and inhibitor design. *Chem. Biol.* **9**, 1085–1094 (2002).
- Weinstein, J. N. et al. An information-intensive approach to the molecular pharmacology of cancer. *Science* **275**, 343–349 (1997).
- Lakey, J. H. & Raggett, E. M. Measuring protein-protein interactions. *Curr. Opin. Struct. Biol.* **8**, 119–123 (1998).
- Gray, N. S. et al. Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors. *Science* **281**, 533–538 (1998).
- Salemme, F. R. Chemical genomics as an emerging paradigm for postgenomic drug discovery. *Pharmacogenomics* **4**, 257–267 (2003).
- MacBeath, G., Koehler, A. N. & Schreiber, S. L. Printing small molecules as microarrays and detecting protein-ligand interactions en masse. *J. Am. Chem. Soc.* **121**, 7967–7968 (1999).
- Winssinger, N., Ficarro, S., Schultz, P. G. & Harris, J. L. Profiling protein function with small molecule microarrays. *Proc. Natl Acad. Sci. USA* **99**, 11139–11144 (2002).
- Falsey, J. R., Renil, M., Park, S., Li, S. & Lam, K. S. Peptide and small molecule microarray for high throughput cell adhesion and functional assays. *Bioconjug. Chem.* **12**, 346–353 (2001).
- Vetter, D. Chemical microarrays, fragment diversity, label-free imaging by plasmon resonance—a chemical genomics approach. *J. Cell Biochem.* **39** (suppl.), 79–84 (2002).
- Birkert, O., Tunnemann, R., Jung, G. & Gauglitz, G. Label-free parallel screening of combinatorial triazine libraries using reflectometric interference spectroscopy. *Anal. Chem.* **74**, 834–840 (2002).
- Birkert, O. & Gauglitz, G. Development of an assay for label-free high-throughput screening of thrombin inhibitors by use of reflectometric interference spectroscopy. *Anal. Bioanal. Chem.* **372**, 141–147 (2002).
- Jona, G. & Snyder, M. Recent developments in analytical and functional protein microarrays. *Curr. Opin. Mol. Ther.* **5**, 271–277 (2003).
- MacBeath, G. Protein microarrays and proteomics. *Nature Genet.* **32** (suppl.), 526–532 (2002).
- Zhu, H. et al. Global analysis of protein activities using proteome chips. *Science* **293**, 2101–2105 (2001).
- Espejo, A., Cote, J., Bednarek, A., Richard, S. & Bedford, M. T. A protein-domain microarray identifies novel protein-protein interactions. *Biochem. J.* **367**, 697–702 (2002).
- Newman, J. R. & Keating, A. E. Comprehensive identification of human bZIP interactions with coiled-coil arrays. *Science* **300**, 2097–2101 (2003).
- Ziauddin, J. & Sabatini, D. M. Microarrays of cells expressing defined cDNAs. *Nature* **411**, 107–110 (2001).

40. Ramachandran, N. *et al.* Self-assembling protein microarrays. *Science* **305**, 86–90 (2004).
41. Lefurgy, S. & Cornish, V. Finding Cinderella after the ball: a three-hybrid approach to drug target identification. *Chem. Biol.* **11**, 151–153 (2004).
42. Liberles, S. D., Diver, S. T., Austin, D. J. & Schreiber, S. L. Inducible gene expression and protein translocation using nontoxic ligands identified by a mammalian three-hybrid screen. *Proc. Natl Acad. Sci. USA* **94**, 7825–7830 (1997).
43. Lunn, M. R. *et al.* Indoprofen upregulates the survival motor neuron protein through a cyclooxygenase-independent mechanism. *Chem. Biol.* **11**, 1495–1503 (2004).
44. Dolma, S., Lessnick, S. L., Hahn, W. C. & Stockwell, B. R. Identification of genotype-selective antitumor agents using synthetic lethal chemical screening in engineered human tumor cells. *Cancer Cell* **3**, 285–296 (2003).
45. Wang, J. & Dreyfuss, G. A cell system with targeted disruption of the SMN gene: functional conservation of the SMN protein and dependence of Gemin2 on SMN. *J. Biol. Chem.* **276**, 9599–9605 (2001).
46. Aiken, C. T., Tobin, A. J. & Schweitzer, E. S. A cell-based screen for drugs to treat Huntington's disease. *Neurobiol. Dis.* **16**, 546–555 (2004).
47. Stegmaier, K. *et al.* Gene expression-based high-throughput screening (GE-HTS) and application to leukaemia differentiation. *Nature Genet.* **36**, 257–263 (2004).
48. Kapur, R. Fluorescence imaging and engineered biosensors: functional and activity-based sensing using high content screening. *Ann. NY Acad. Sci.* **961**, 196–197 (2002).
49. Yarrow, J. C., Perlman, Z. E., Westwood, N. J. & Mitchison, T. J. A high-throughput cell migration assay using scratch wound healing, a comparison of image-based readout methods. *BMC Biotechnol.* **4**, 21 (2004).
50. Kau, T. R. *et al.* A chemical genetic screen identifies inhibitors of regulated nuclear export of a Forkhead transcription factor in PTEN-deficient tumor cells. *Cancer Cell* **4**, 463–476 (2003).
51. Root, D. E., Flaherty, S. P., Kelley, B. P. & Stockwell, B. R. Biological mechanism profiling using an annotated compound library. *Chem. Biol.* **10**, 881–892 (2003).
52. Seidler, J., McGovern, S. L., Doman, T. N. & Shoichet, B. K. Identification and prediction of promiscuous aggregating inhibitors among known drugs. *J. Med. Chem.* **46**, 4477–4486 (2003).
53. Tuschl, T. Expanding small RNA interference. *Nature Biotechnol.* **20**, 446–448 (2002).
54. Elbashir, S. M. *et al.* Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411**, 494–498 (2001).
55. Lassus, P., Rodriguez, J. & Lazebnik, Y. Confirming specificity of RNAi in mammalian cells. *Sci. STKE* **147**, PL13 (2002).
56. Root, D. E., Kelley, B. P. & Stockwell, B. R. Global analysis of large-scale chemical and biological experiments. *Curr. Opin. Drug Discov. Dev.* **5**, 355–360 (2002).
57. Burke, T. J., Loniello, K. R., Beebe, J. A. & Ervin, K. M. Development and application of fluorescence polarization assays in drug discovery. *Comb. Chem. High Throughput Screen.* **6**, 183–194 (2003).
58. Timasheff, S. N., Andreu, J. M. & Na, G. C. Physical and spectroscopic methods for the evaluation of the interactions of antimitotic agents with tubulin. *Pharmacol. Ther.* **52**, 191–210 (1991).
59. Bulseco, D. A. & Wolf, D. E. Fluorescence correlation spectroscopy: molecular complexing in solution and in living cells. *Methods Cell Biol.* **72**, 465–498 (2003).
60. Misra, R. Modern drug development from traditional medicinal plants using radioligand receptor-binding assays. *Med. Res. Rev.* **18**, 383–402 (1998).
61. Hicks, R. P. Recent advances in NMR: expanding its role in rational drug design. *Curr. Med. Chem.* **8**, 627–650 (2001).
62. Siegel, M. M. Early discovery drug screening using mass spectrometry. *Curr. Top. Med. Chem.* **2**, 13–33 (2002).
63. Homola, J. Present and future of surface plasmon resonance biosensors. *Anal. Bioanal. Chem.* **377**, 528–539 (2003).
64. Jelezarov, I. & Bosshard, H. R. Isothermal titration calorimetry and differential scanning calorimetry as complementary tools to investigate the energetics of biomolecular recognition. *J. Mol. Recogn.* **12**, 3–18 (1999).
65. Burke, M. D., Berger, E. M. & Schreiber, S. L. Generating diverse skeletons of small molecules combinatorially. *Science* **302**, 613–618 (2003).
66. Oprea, T. I. & Matter, H. Integrating virtual screening in lead discovery. *Curr. Opin. Chem. Biol.* **8**, 349–358 (2004).
67. Ewing, T. J., Makino, S., Skillman, A. G. & Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des.* **15**, 411–428 (2001).
68. Osterberg, F., Morris, G. M., Sanner, M. F., Olson, A. J. & Goodsell, D. S. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins* **46**, 34–40 (2002).
69. Kramer, B., Rarey, M. & Lengauer, T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins* **37**, 228–241 (1999).
70. Halgren, T. A. *et al.* Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **47**, 1750–1759 (2004).
71. Vangrevelinghe, E. *et al.* Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. *J. Med. Chem.* **46**, 2656–2662 (2003).
72. Peng, H. *et al.* Identification of novel inhibitors of BCR-ABL tyrosine kinase via virtual screening. *Bioorg. Med. Chem. Lett.* **13**, 3693–3699 (2003).
73. Bajorath, J. Integration of virtual and high-throughput screening. *Nature Rev. Drug Discov.* **1**, 882–894 (2002).
74. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **46**, 3–26 (2001).
75. Hann, M. M. & Oprea, T. I. Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* **8**, 255–263 (2004).

Acknowledgements B.R.S. is supported in part by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund.

Competing interests statement The author declares that he has no competing financial interests.

Navigating chemical space for biology and medicine

Christopher Lipinski¹ & Andrew Hopkins²

¹Pfizer Global R&D, Groton Laboratories, Eastern Point Road, Groton, Connecticut 06340, USA (e-mail: christopher_a_lipinski@groton.pfizer.com)

²Pfizer Global R&D, Sandwich Laboratories, Ramsgate Road, Sandwich, Kent CT13 9NJ, UK (e-mail: andrew.hopkins@pfizer.com)

Despite over a century of applying organic synthesis to the search for drugs, we are still far from even a cursory examination of the vast number of possible small molecules that could be created. Indeed, a thorough examination of all 'chemical space' is practically impossible. Given this, what are the best strategies for identifying small molecules that modulate biological targets? And how might such strategies differ, depending on whether the primary goal is to understand biological systems or to develop potential drugs?

The relationship between chemistry, biology and medicine has been remarkably productive over the past century, since Paul Ehrlich pioneered the idea of systematically searching for drugs. By screening just over 600 synthetic compounds, Ehrlich discovered arsphenamine (Salvarsan)¹, which greatly improved the treatment of syphilis. Researchers now routinely screen millions of compounds in the search for some that are biologically active. Yet even the compound files of the largest pharmaceutical companies (which typically contain approximately 10^6 compounds) offer only a cursory

examination of all the possible organic compounds that comprise 'chemical space' (Box 1). Chemical space is for all practical purposes infinite and limited only by the chemist's imagination.

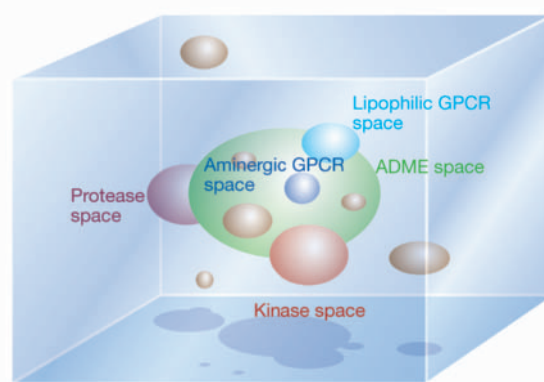
Not all biologically active compounds have the desired physicochemical properties to be a drug. A biologically active compound may be too lipophilic (greasy) to be orally absorbed, too polar to cross the gastrointestinal wall or may have too much vulnerable chemical functionality that can be attacked by metabolizing systems in the liver, and therefore not remain intact for long enough to have a useful *in vivo* effect.

Box 1

Chemical space

Chemical space can be viewed as being analogous to the cosmological universe in its vastness, with chemical compounds populating space instead of stars. For example, there are more than 10^{29} possible derivatives of *n*-hexane — if we use a list of only 150 substituents and consider mono- to 14-substituted hexanes⁵⁰. However, not all theoretically postulated compounds fall within the limits of what is synthetically feasible to produce, even with our current, extensive knowledge of organic chemistry. To navigate the vast diversity of chemical space, the concept of 'chemography', which is akin to a global positioning system, has been proposed. This involves mapping compounds onto coordinates of chemical descriptors of various physicochemical or topological properties^{51,52}. Given the vastness of chemical space, the challenge for chemical biologists and drug discoverers is to identify those regions that are likely to contain biologically active compounds, that is, biologically relevant chemical space. The limits of biologically relevant chemical space are defined by the specific binding interactions between small molecules and the three-dimensional molecular recognition patterns on biological molecules, such as proteins, RNA and DNA, which have evolved over billions of years.

Measured in terms of physicochemical properties and topological descriptors, therapeutically useful compounds appear to cluster together in galaxies. A major unknown is whether these galaxies are evenly and sparsely distributed and therefore hard to find, or whether most of the chemical universe is 'empty' (containing no therapeutically interesting compounds), with galaxies of therapeutically interesting compounds scattered far apart. A century of medicinal chemistry and thousands of high-throughput screening programmes suggests that compounds that bind to certain



Box 1 Figure The figure depicts a cartoon representation of the relationship between the continuum of chemical space (light blue) and the discrete areas of chemical space that are occupied by compounds with specific affinity for biological molecules. Examples of such molecules are those from major gene families (shown in brown, with specific gene families colour-coded as proteases (purple), lipophilic GPCRs (blue) and kinases (red)). The independent intersection of compounds with drug-like properties, that is those in a region of chemical space defined by the possession of absorption, distribution, metabolism and excretion properties consistent with orally administered drugs — ADME space — is shown in green (see Box 2).

'target classes' (proteins from the same superfamily, such as G-protein-coupled receptors; GPCRs) are clustered together in discrete regions of chemical space (see figure). These regions can be defined by particular chemical descriptors.

Box 2

What do drugs look like?

Drug-likeness

The distribution of the molecular properties of small-molecule launched drugs has changed little in the past 20 years, despite changes in the types of clinical indication for which drugs have been discovered and the range of targets acted upon⁵³. Lipinski's seminal analysis of the Derwent World Drug Index introduced the concept of drug-likeness: orally administered drugs are far more likely to reside in areas of chemical space defined by a limited range of molecular properties. These properties have been encapsulated in Lipinski's 'rule of five'. This analysis shows that, historically, 90% of orally absorbed drugs had fewer than five hydrogen-bond donors, less than ten hydrogen-bond acceptors, molecular masses of less than 500 daltons and log P values (a measure of lipophilicity) of less than five². Since this work, various definitions of, and methods to predict, drug-likeness have been proposed^{2,54–65}. However, the consensus is that drug-likeness is defined

by a range of molecular properties and descriptors that can discriminate between drugs and non-drugs for such characteristics as oral absorption, aqueous solubility and permeability. Computational property filters can be used to rapidly assess the drug-likeness of chemical libraries *in silico* before purchase or synthesis²¹.

Druggability

The concept of druggability postulates that since the binding sites on biological molecules are complementary with their ligands in terms of volume, topology and physicochemical properties, then only certain binding sites on putative drug targets will be compatible with high-affinity binding to compounds with drug-like properties³¹. The extension of this concept to a whole genome analysis leads to the identification of the druggable genome. This is the expressed proteome predicted to be amenable to modulation by compounds with drug-like properties³⁰.

Recently, toxicity has replaced poor drug metabolism properties as a major cause of failure in the early clinical phase of drug discovery.

The determination of the characteristics of compounds that are more likely to yield safe, orally bioavailable medicines has led to the concept of 'drug-likeness'. Compounds that are drug-like have the potential to be developed into orally administered drugs (Box 2; ref. 2), which are generally favoured owing to their ease of use by patients. But biologically active compounds that do not have the exacting properties required of a drug can nevertheless be extremely useful to

science as 'tools' for dissecting biological mechanisms and testing hypotheses in model systems. In recent years, it has been argued that it would be useful to discover a chemical tool to modulate every known protein³. Indeed, the Molecular Libraries Screening Center Network that is being established as part of the recent National Institutes of Health (NIH) Roadmap is aiming to facilitate the discovery of new chemical tools to understand biology, some of which may aid future drug development⁴. This Roadmap will allow the public sector to obtain data from high-throughput screens of a large

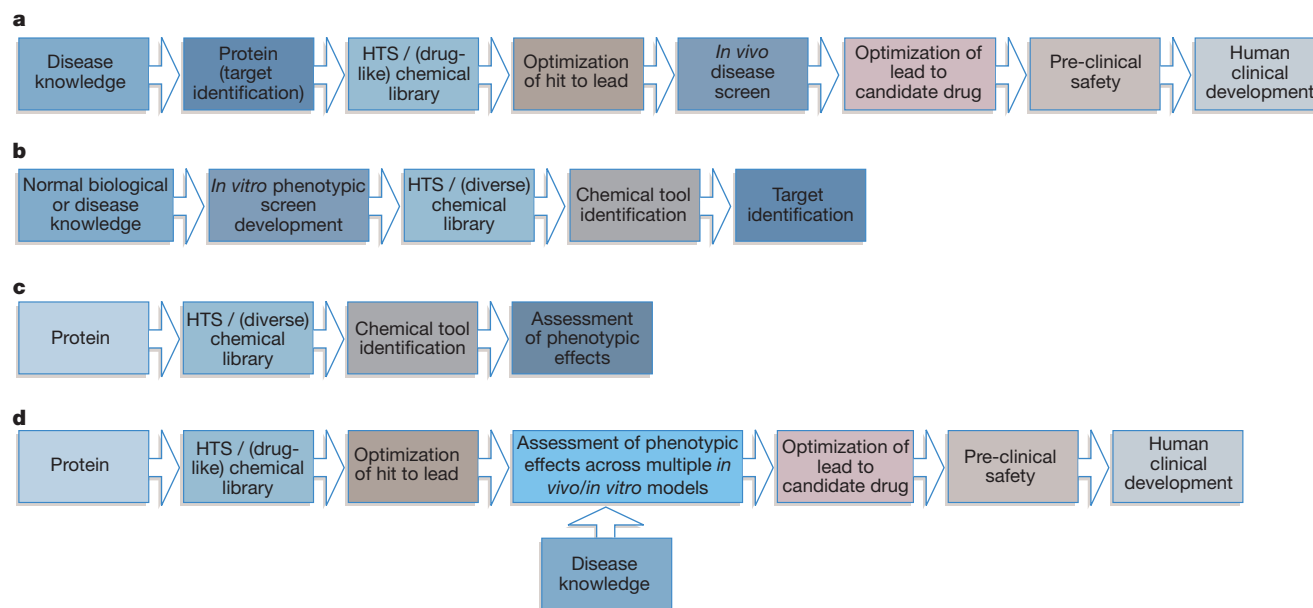


Figure 1 A comparison of approaches to discovering small-molecule tools or drugs. **a**, The 'standard model' of drug discovery is considered to be a linear process. New targets (usually proteins) are identified through knowledge of a particular disease. Compounds in drug-like (see Box 2) chemical libraries are tested in high-throughput screens (HTS) for their ability to bind to or modulate the target of interest. Selected initial hits (compounds that show levels of activity beyond a certain threshold level in the screen) are subsequently optimized through testing in further screens (often lower throughput) to give leads that have the required pharmacokinetic properties. These are then tested *in vivo*. Leads showing the required efficacy in *in vivo* disease models are further optimized into clinical drug candidates, which are then tested in human clinical trials. **b**, By comparison, forward chemical genetic approaches to developing chemical tools often start by

screening a diverse chemical library to identify chemical tools that induce a particular phenotypic effect (such as cell death or cell proliferation). In phenotypic screens, the specific target of the chemical tools is often unknown, so a subsequent stage of target identification is required. **c**, Reverse chemical genetic approaches begin with a target of interest and then attempt to discover a specific chemical tool that binds to the target, usually by screening a diverse chemical library against the target *in vitro*. The specific chemical tool is then assessed for its ability to cause a range of phenotypic effects to identify the function of the target. **d**, Combining chemical tool and drug discovery approaches can result in an alternative drug discovery strategy to the standard model. Here, specific chemical tools are screened empirically across several disease models to discover new therapeutic effects.

Table 1 Rough guide to some major *in vivo* biological tools

Biological Tool	Description	Advantages	Disadvantages	Reference
Protein-specific antibodies	Immunoglobulins secreted from a single clone of antibody-producing cells can be used as immunochemical reagents for the detection or assay of particular antigens	- Highly specific for intended target - Can be generated for virtually any species with known immunoglobulin genes	- Specifically target only circulating and cell-surface proteins <i>in vivo</i> - Lack of cross-species selectivity with some targets	34
Engineered recombinant proteins	A technique of subtly modifying endogenous proteins to produce ligands with agonist or antagonist functions	Single amino-acid changes are often sufficient to produce antagonist ligands from naturally secreted proteins	- Predominantly target circulating and cell-surface proteins <i>in vivo</i> - Lack of cross-species selectivity with some targets	35
Gene knockouts	A technique in which homologous recombination is used to inactivate a gene from its endogenous locus to create a null phenotype	Applicable to routine mass production of gene knockouts in mice	Irreversible gene knockouts are not capable of mimicking the temporal and dynamic nature of perturbation that is characteristic of chemical or protein tools	36
Gene knockins	A technique in which homologous recombination is used to insert a gene into a selected chromosomal location	- Engineered site-directed mutant knockins can be used to design selective chemical-tool-receptor systems - Can be used to induce gain-of-function mutants to mimic agonist activity	- Initial chemical tool and/or structural knowledge of binding site required for site-directed engineering - Gain-of-function mutants are permanent, so do not mimic the dynamic nature of agonism	37
RNA interference	A phenomenon in which the expression of a gene is temporarily inhibited when a double-stranded complementary RNA is introduced into the organism	- Targeted libraries can be produced on a genome-wide scale for mammalian cell systems - Mass-production <i>in vivo</i> RNAi methods available for worm, fly and zebrafish	- Introduction of foreign double-stranded RNA into cells often results in an interferon response <i>in vivo</i> and other off-target effects - Genes with high messenger RNA turnover/low protein turnover can be difficult to target - Chronic <i>in vivo</i> administration difficult owing to delivery by large-volume hydrodynamic intravenous injections	38
Intrabodies	Intracellular antibodies composed of single-chain Fv fragments that consist of heavy- and light-chain antibody variable domains	- Target previously inaccessible intracellular proteins - Act at protein level, interacting directly with target of interest - Potential to act as agonists as well as antagonists	Whole-animal <i>in vivo</i> delivery using viral vectors or liposomes is currently problematic	39

collection of compounds (initially about 500,000 compounds) in various biological assays. Here, we consider the scientific and practical issues that need to be addressed if efforts to discover new chemical tools are to provide the maximum possible benefit.

Chemical tools versus biological tools

Before the molecular biology revolution, the tools of the pharmacologist were usually the only ones available for probing the behaviour of biological systems. The pharmacologist's tools were mostly chemicals, derived from natural sources or from chemical synthesis. Perturbations of biological systems using such tools, some of which led to the development of drugs, taught us much about biology. For example, the natural product staurosporine — used as an early tool to probe the effects of tyrosine kinase inhibition — was important in the discovery of the anticancer drug imatinib (Gleevec), an inhibitor of the BCR-ABL tyrosine kinase.

However, the discovery of a new pharmacological tool was, and still is, a relatively rare and somewhat serendipitous event. At the core of efforts to discover small molecules of biological interest is typically some form of biological screen, in which a collection of compounds (known as a library) is assayed for a particular biological activity. In the early era of pharmacology, the compounds were often derived from natural sources, and the assays were for effects such as anti-bacterial activity or anti-inflammatory activity, usually using *in vivo* primary screens. More recently, with the molecular biology revolution, screening against isolated macromolecular targets has become widespread, and the compounds screened are often purely synthetic products from combinatorial chemistry (an approach for creating molecules *en masse*) as opposed to natural products⁵. Indeed, since the publication of the first paper to describe the synthesis of a single combinatorial library in 1992 (ref. 6), there has been a considerable increase in the numbers of combinatorial-chemistry compounds

being developed for high-throughput screening experiments. For example, the sixth annual Comprehensive Survey of Combinatorial Library Synthesis records a total of 338 chemical libraries published in 2002 — a 25% increase from the previous year⁷. The success of combinatorial chemistry so far is hard to assess because of the 10- to 15-year time lag between initial chemical synthesis and drug launch. However, the general consensus is that many of the compounds made in the early years of combinatorial chemistry (from about 1992 to 1997) were severely flawed. Growing appreciation of the underlying reasons for this has considerably improved current combinatorial chemistry. The lesson to be learned here is that a radically new technology such as combinatorial chemistry may take well over a decade to mature and become fully useful.

Screening of small molecules is still the technology of choice for the development of many human medicines (Fig. 1), owing to its compatibility with the production of orally administered drugs. But for investigating biological function, biological tools have been in ascendance. These are created by genetic and protein engineering techniques, and are both cheaper and more efficient to develop than small-molecule chemical tools. Over the past decade, biologists wishing to probe protein function have invented an ever-growing array of techniques to manipulate and perturb biological systems (see Table 1).

It is estimated that the databases of the world pharmaceutical companies collectively contain small-molecule compounds known to directly modulate the function of only around 1,000 proteins (although few small-molecule compounds would be considered wholly selective)⁸. However, using genetic techniques, biologists can now readily selectively delete or silence the expression of almost any gene in the genomes of several diverse model organisms, including yeast, worm, fruitfly, zebrafish and mouse. Such genetic methods to explore the function of specific genes are on average 10 to 1,000 times

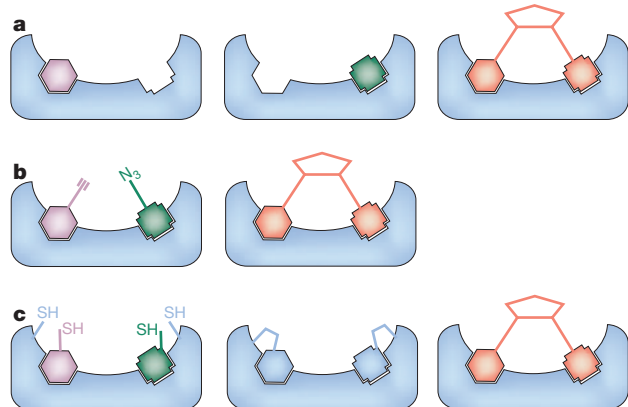


Figure 2 Fragment-based lead screening. Methods are currently being developed to more effectively search chemical space by screening a relatively small number of low-molecular-weight 'fragments'. Although in theory an unimaginably large number of drug-like compounds can be synthesized, chemical space grows as a function of the number of atoms in the compound (that is, its molecular weight): as the number of atoms increases, so does the number of possible permutations. So, restricting the molecular weight by screening only fragments and scaffolds allows a much larger area of possible chemical space (for a given number of atoms) to be explored. Low-molecular-weight, weak-binding fragments are then connected to form high-affinity, higher-molecular-weight ligands⁴⁰. Several approaches based on this idea have been developed. **a**, Nuclear magnetic resonance (NMR)⁴¹, mass spectrometry⁴² and X-ray crystallography^{43–45} are used to screen for low-affinity fragments (shown in mauve and green). Information on the structure–activity relationships (SAR) from these approaches can be used to rationally link fragments that bind in different parts of the target binding site to give larger, high-affinity ligands (shown in orange). **b**, The binding site can also be used to 'guide' the self-assembly of fragments (shown in mauve and green) containing chemical groups that can react to link the fragments to give high-affinity ligands (shown in orange) — an approach known as 'click chemistry'^{46,47}. **c**, An approach known as 'tethering' can also be used to identify fragments (shown in mauve and green) that form covalent disulphide bonds to engineered cysteine residues within the binding site^{48,49}; again, these fragments can be combined to produce larger, high-affinity ligands (shown in orange). Approaches such as tethering can also help to identify small molecules that bind to protein targets in cases where high-throughput screening approaches have been unsuccessful (for example, the so-called 'undruggable' targets).

less expensive than current chemical-based methods (R. W. Spencer, personal communication). This is exemplified by the fact that it is possible for a small biotechnology company to produce knockout mice for every member of the 'druggable genome' (Box 2) in only a few years⁹. Even with the combined screening resources of the top ten pharmaceutical companies, several years and several billion dollars would be required to produce the equivalent number of chemical tools from screening for the same set of targets. This is illustrated by the following rough calculation: it is not uncommon in industry to screen one million compounds per high-throughput screening campaign; so, if the total screening cost was as low as US\$0.4 per compound (R. W. Spencer, personal communication), including the cost of the chemical synthesis, high-throughput-screening disposables, capital costs and human resources, screening just 25 targets with one million compounds would cost US\$10 million, and screening the estimated 2,500 druggable targets in the human proteome would cost approximately US\$1 billion.

Despite advances in the development of biological tools, many such tools have severe limitations, particularly when it comes to investigating the dynamic, reversible and temporal elements of protein function. In addition, although biological tools can antagonize the function of a protein by preventing or reducing its

expression, or by blocking its ability to bind to other proteins, few biological tools allow the mimicking of 'agonist' behaviour by causing gain-of-function. This can be achieved more readily in certain gene families, such as G-protein-coupled receptors, by using small molecules. The realization of these limits has resulted in a revival of traditional small-molecule approaches to understanding biological function. Such approaches — now re-branded as 'chemical genetics' or 'chemical genomics' — are similar in character to the empirical investigational methods of pharmacology and physiology (see Fig. 1)^{3,10}.

Chemical tools are also important for target validation in drug discovery; that is, they can be used to verify whether a protein is a suitable target for drug development. One assumption underlying the chemical genetics approach is that the chemical tools are sufficiently selective in their modulating activity that an altered function can be assigned to a specific protein. However, compared with the exquisite selectivity of many biological techniques, selectivity cannot be commonly assumed for small molecules^{11,12}. Moreover, although our knowledge of the desired properties of chemicals intended to become drugs is growing, very little is known about the chemical characteristics required of tools when the goal is something other than drug discovery.

Drug discovery versus knowledge discovery

The desired properties of chemical tools in the broadest sense depend in part on the goal of the experimenter: chemical geneticists aim to use small molecules to explore biological function; those involved in drug discovery and development aim to find small molecules that achieve a desired therapeutic effect in humans without causing unacceptable side effects. The importance of this difference in goals can be appreciated by briefly contrasting the general approach and priorities of the pharmaceutical industry with the approach and priorities of academic laboratories involved in chemical genetics.

The current primary strategy of the pharmaceutical industry for identifying biologically active molecules that might be starting points for potential drugs is the use of high-throughput screening. Here, libraries of about 10^5 to 10^6 small molecules with some drug-like characteristics are screened in high-throughput assays. These assays measure the ability of the small molecules to modulate a particular biological target, and vast amounts of data are generated. However, what is perhaps not widely appreciated by those outside industry is the generally poor quality of these data. For example, when an identical set of compounds is screened against the same biological target using three different assay formats, the concordance in the number of biologically active compounds or 'hits' obtained from each assay is just 35%. This is due in part to the inherent noise in the assays^{13,14}, although reproducibility within each individual assay is much more robust. Nevertheless, this low quality is acceptable to industry, as long as some active compounds are identified that have the potential to be optimized using more rigorous, lower-throughput assays. In other words, the high-throughput-screening process merely serves as a coarse 'filter' on the route to locating a potential drug; the limited number of positive hits are used to direct further experiments. The 'negative' information is too coarse to falsify hypotheses such as whether a particular type of chemical structure does not have a particular effect, but it can be exploited to identify borderline hits by computational pattern recognition¹⁵ and probabilistic data mining¹⁶. Several factors contribute to the limitations of negative data, such as the fact that, with rare exceptions, compound concentrations are unknown in high-throughput screening because of well-documented compound solubility problems, both in dimethylsulphoxide (DMSO) stocks and upon dilution with aqueous buffer¹⁷. A compound may therefore appear inactive because it is truly inactive, or simply because its concentration was much lower than that assumed.

By contrast, in chemical genetics studies carried out in academic laboratories, collections of small molecules are typically screened in

assays for their effects on processes such as cell death, cell migration and cell proliferation. A key aim of these studies is to identify correlations between different experiments that will aid in understanding the basis of the biological activities observed. Such experiments are discussed extensively in the review in this issue by Stockwell (page 846); but the key point to appreciate here is that data quality for such experiments is crucial. Given this, what approaches might be the most appropriate for identifying new chemical tools?

Searching for the right chemical tool

Suppose that the goal is to interrogate a biological system with a small molecule and that we restrict ourselves to using only robust, positive information that has survived a filtering process of experimental re-testing. Furthermore, let us agree that we want to generate useful information in a tool sense; that is, our aim is to learn something about biological function, whether or not it has any relevance to human therapeutics. What properties does the tool need to have? Must we restrict ourselves to using drug-like compounds? Two chemical extremes can be discerned: tools with properties consistent with their development into oral drugs; and tools with properties that could confound their development into oral drugs. Of course, many compounds will lie between these extremes. By discussing both these extremes here, we attempt to illustrate the considerations that could be important for initiatives aimed at developing chemical tools to explore biology and/or to act as a starting point for drug development.

Tools with drug-like properties

At one extreme, the chemical nature of the tool itself is drug-like, although the tool does not necessarily have all the attributes required of a drug (see Box 2). The main advantage here is that, should modulating the target of an identified tool compound be of therapeutic interest, this tool compound will be a suitable starting point for drug development. Another advantage is that limiting the search for tools to drug-like compounds means avoiding the potential pitfalls associated with compounds that contain chemical groups associated with toxic effects, or compounds that interact covalently with protein targets. The latter suffer from problems such as lack of specificity and unsuitability for optimization by medicinal chemistry techniques^{18,19}. Although several well-known drugs, such as omeprazole and β -lactamase inhibitors, are known to act by means of irreversible mechanisms²⁰, medicinal chemists and toxicologists are becoming more wary of incorporating reactive groups within tools or drugs that can form covalent bonds to the target and/or other proteins. A disadvantage of drug-like libraries is that the breadth of commercially available chemistry space is decreased by the order of 50% to 80% (ref. 21). Another disadvantage is cost; frequently, drug-like compounds are more expensive to purchase than non-drug-like compounds.

Tools that are not like oral drugs

At the other extreme, the tool is not drug-like; chemical 'flaws' are present that mean the compound is unlikely to be used to treat human disease. For example, a moiety associated with toxicity can be present in a tool, provided that the unwanted toxicity does not present itself in the timescale of experiments using the tool, or if the tool is only intended for use in systems where toxicity is not an issue. An advantage of this type of tool is that the commercially available chemistry space is larger. Another very considerable advantage is that the interrogation of biology is unhindered by other drug discovery considerations, such as the need for the tool compound to be orally bioavailable.

Nevertheless, chemical genetics requires selective tools to interrogate and dissect biological processes. Lack of selectivity in a tool with 'relaxed' chemistry criteria (that is, chemical structural features known to be associated with increased probability of drug discovery failure) is a very real possibility; chemical features associated with failure in drug discovery tend to cause compounds to have 'promiscu-

ous' effects in biological systems. A clear example of this would be the presence of a functional group that is likely to interact covalently with proteins (such as an epoxide or an aldehyde) in a simple, featureless, low-complexity compound (the complexity of a compound is related to the character and number of functional groups within the compound; see also ref. 22). This is because a low-complexity compound has a higher probability of weak binding to a target and a higher probability of binding to many targets²³. When a low-complexity compound irreversibly binds to many targets (for example, several proteins) by means of covalent chemical bonds, the complexity of the biological effects elicited is very large. Thus, deciphering the effects of the compound as a tool is difficult. Whereas biological tools can be designed to be exquisitely selective for a particular gene or protein, it is harder to make the same selectivity prediction for any small-molecule chemical tool^{12,13}.

Another disadvantage of chemical tools that are not drug-like is the lack of clarity as to whether chemical features will defeat the utility of the tool. The available chemical space is likely to increase as chemistry criteria are relaxed. But if the aim is to use the chemical tool in *in vivo* animal models, which may have more relevance to both normal biology and disease than *in vitro* systems, then consideration of the drug-like properties of the tool, in terms of pharmacokinetics and the therapeutic index between efficacy and toxicity, is vital.

Relationship between tools and models

Whether the aim is to discover drugs or to gain knowledge of biological systems, the nature and properties of a chemical tool cannot be considered independently of the system it is to be tested in. Compounds that bind to isolated recombinant proteins are one thing; chemical tools that can perturb cell function another; and pharmacological agents that can be tolerated by a live organism and perturb its systems are yet another. If it were simple to ascertain the properties required to develop a lead discovered *in vitro* to one that is active *in vivo*, drug discovery would be as reliable as drug manufacturing. Indeed, examples abound of experimental drugs with the same primary effect in an isolated *in vitro* assay (such as antagonism of a particular protein) failing in clinical development because of inappropriate pharmacokinetics and/or toxicity. For example, the first histamine H2 receptor antagonist to be tested clinically was burimamide. Its pharmacokinetic properties were not compatible with oral administration, but tested parenterally (administered in a manner other than through the digestive tract), it was used to prove that inhibiting histamine H2 receptors effectively inhibited gastric-acid secretion. Metiamide, the second H2 receptor antagonist tested in humans, was orally active, but clinical trials were terminated because it caused fatal bone marrow toxicity. Cimetidine was the third H2 receptor antagonist to reach the clinic. This orally active compound was devoid of the toxicity found in metiamide and became the world's first billion-dollar blockbuster drug; its safety is attested to by its eventual over-the-counter availability worldwide. In general (whether we are considering either tools or potential drugs) because of the uncertainty of whether a compound has all the required properties to act effectively at a specific point in a whole organism, we cannot falsify a hypothesis about the biological function of a specific protein unless dosing effects, pharmacokinetics and selectivity are understood. All this requires significant investment and investigation.

If our goal is to discover chemical tools that bind to isolated recombinant proteins, then several emerging chemical technologies based on screening low-molecular-weight chemical 'fragments' may allow a more effective exploration of chemical space than the high-throughput screening of large chemical libraries (new approaches in this area are discussed in Fig. 2 and Box 3). However, the 'reductionist' approach of screening for small-molecule hits in isolated assay systems that bear little resemblance to the biological systems in which they are meant to act may be partly responsible for the decline in drug discovery productivity of the pharmaceutical industry over the past decade^{24,25}.

Box 3

Protein dynamics and chemical space

The experience of the pharmaceutical industry in screening thousands of protein targets indicates that not all proteins are amenable to small-molecule modulation. Those that are not are called undruggable targets. A great deal of investment can therefore prove fruitless. Nevertheless, occasionally an unexpected allosteric binding site for a drug or chemical tool is discovered. Therefore, methods to discover ligands for unpredicted binding sites could improve the cost-effectiveness of searches for chemical tools. A particular challenge is to identify *a priori* which undruggable proteins are flexible enough to accommodate allosteric binding sites. A more effective method might be to combine several technologies. For example, the computational ability to scan protein structures *in silico* for flexible 'hot spots' — protein features that are likely to interact with small organic molecules^{32,66} — could be coupled with protein binding site analysis³¹ and substantial improvements in 'docking and scoring'. The latter involves the computational prediction of the binding of small-molecule ligands to the structure of a protein derived from an experimental X-ray or NMR structure.

The reductionist approach is a powerful aid to medicinal chemistry, in terms of understanding the potency and selectivity of small molecules for particular isolated protein targets. However, it has been criticized for its over-emphasis on potency and selectivity; pharmacokinetic properties, toxicity and biological responses have, in general, been beyond our ability to model or predict. Indeed, in drug discovery, despite the existence of stringent drug-like criteria, drugs and leads are rarely wholly selective for one target. In fact, 'poly-pharmacology' is often the basis for a drug's efficacy (as illustrated by many psychiatric compounds and a growing realization of the role of promiscuity in the generation of new kinase inhibitors in oncology²⁶).

The most efficient strategy for finding chemical tools or drugs is often mistaken for the most effective strategy. Although the modern reductionist *in vitro* methods used in early discovery are efficient at discovering hits against new, isolated targets, it seems more difficult to convert such hits into drugs. This might also be a challenge for those trying to develop effective tools for probing *in vivo* biology. However, historically, the problem of the *in vivo* screening approach has been the difficulty in discovering new leads for new targets. The use of phenotypic *in vivo* screens as primary assays requires a compound to have suitable absorption, solubility and permeability characteristics, in addition to high potency at a given target and relatively low toxicity, for activity to be detected. Nevertheless, as our understanding of the properties of drugs, as distinct from the rest of the chemical universe, increases (see Box 2), lessons from the historical approach are being applied to high-throughput screening assays and chemical-library design for drug discovery²⁷. Such knowledge could also be valuable in the development of chemical tools.

Perspective

Is the goal of discovering a small molecule to modulate every known protein practical? Decades of medicinal chemistry experience within the pharmaceutical industry suggest otherwise. Although less-stringent criteria applied to chemical tools can expand the accessible biological target space, toxicological and pharmacokinetic considerations must be taken into account if the use of a tool is to extend beyond isolated protein-binding assays to probing whole animal systems. So, for some targets, it may be more cost-effective to generate biological tools.

Some argue that the only limit to developing a chemical tool for a given protein target is the diversity of the chemicals screened against it. However, the evolution of specific molecular recognition by proteins creates stringent physicochemical limits that restrict the target set available to modulation by small molecules. These con-

straints are more severe if the aim is to discover drugs that can be orally administered. Furthermore, while many pharmaceutical companies and academics have been following the assumption that using a larger array of diverse chemistry to explore wider areas of chemical space will lead to the discovery of new drugs, the most successful drug discoverer to date, the late Paul Janssen, was surprisingly conservative in his exploration of chemical space²⁸. Janssen's drugs show a steady evolution in their structures because he understood the constraints of biological activity, pharmacokinetics and toxicology on chemical space. Janssen's conservatism in chemical space provided an anchor from which he could be more creative in exploring indications (uses) of a drug through clinical experiments and observations. It is this multiplicity of constraints and competing demands on drug discovery that led another great drug hunter, Sir James Black, to advise: "the most fruitful basis for the discovery of a new drug is to start with an old drug" (refs 8, 29).

Given our limited resources, is the best strategy to explore as much of the diversity of chemical space as possible, or to focus our explorations? In our view, a concerted effort by industry and academia to develop chemical tools to modulate those proteins that make up the predicted druggable subset of the proteome³⁰, and to make these available to all researchers, along with relevant pharmacokinetic data (for *in vivo* models), is an achievable goal that would be of great benefit to biological and medical research. We believe that this goal should be prioritized before resources are expended in the search for chemical tools to modulate proteins that are inherently less tractable to this approach (Box 3). Examples of such proteins are the many proteins that participate in protein-protein interactions in biological signalling cascades^{31,32}. Alternative approaches, such as the use of monoclonal antibodies, might well be more cost-effective for such targets.

The discovery of new pharmacological tools may depend on the serendipity of screening until more effective design methods are devised (Fig. 2). Ultimately, our explorations of biologically relevant chemical space are not limited by our chemical imagination, but by the limits of protein architecture and flexibility (Box 3)³³. Improving our ability to discover new chemical tools and medicines will require combining the efficiency of exploration gained by reductionism with the effectiveness of approaches that study biological systems as a whole. □

doi:10.1038/nature03193

1. Sneader, W. *Drug Prototypes and their Exploitation* (Wiley, London, 1996).
2. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **23**, 3–25 (1997).
3. Schreiber, S. L. Chemical genetics resulting from a passion for synthetic organic chemistry. *Bioorg. Med. Chem.* **6**, 1127–1152 (1998).
4. Austin, C. P., Brady, L. S., Insel, T. R. & Collins, F. S. NIH molecular libraries initiative *Science* **306**, 1138–1139 (2004).
5. Bleicher, K. H., Böhm, H. J., Müller, K. & Alanine, A. I. Hit and lead generation: beyond high-throughput screening. *Nature Rev. Drug Discov.* **2**, 369–378 (2003).
6. Bunin, B. A. & Ellman, J. A. A general and expedient method for the solid-phase synthesis of 1,4-benzodiazepine derivatives. *J. Am. Chem. Soc.* **114**, 10997–10998 (1992).
7. Dolle, R. E. Comprehensive survey of combinatorial library synthesis: 2002. *J. Comb. Chem.* **5**, 693–753 (2003).
8. Wermuth, C. G. Selective optimization of side activities: another way or drug discovery. *J. Med. Chem.* **47**, 1303–1314 (2004).
9. Zambrowicz, B. P. & Sands, A. T. Knockouts model the 100 best-selling drugs — will they model the next 100? *Nature Rev. Drug Discov.* **2**, 38–51 (2003).
10. Crews, C. M. & Splittgerber, U. Chemical genetics: exploring and controlling cellular processes with chemical probes. *Trends Biochem. Sci.* **24**, 317–320 (1999).
11. McGovern, S. L. & Shoichet, B. K. Kinase inhibitors: not just for kinases anymore. *J. Med. Chem.* **46**, 1478–1483 (2003).
12. Krejsa, C. M. *et al.* Predicting ADME properties and side effects: the BioPrint approach. *Curr. Opin. Drug Discov. Devel.* **6**, 470–480 (2003).
13. Wu, X., Glickman, J. F., Bowen, B. R. & Sills, M. A. Comparison of assay technologies for a nuclear receptor assay screen reveals differences in the sets of identified functional antagonists. *J. Biomol. Screen.* **8**, 381–392 (2003).
14. Sills, M. A. *et al.* Comparison of assay technologies for a tyrosine kinase assay generates different results in high throughput screening. *J. Biomol. Screen.* **7**, 191–214. (2002).
15. Harper, G., Bradshaw, J., Gittins, J. C., Green, D. V. & Leach, A. R. Prediction of biological activity for high-throughput screening using binary kernel discrimination. *J. Chem. Inf. Comput. Sci.* **41**, 1295–1300 (2001).

16. Engels, M. F. M., Wouters, L., Verbeeck, R. & Vanhoof, G. Outlier mining in high throughput screening experiments. *J. Biomol. Screen.* **7**, 341–351 (2002).
17. Lipinski, C. A. in *Methods and Principles in Medicinal Chemistry* (eds van de Waterbeemd, H. *et al.*) **18**, Ch. 9 215–231 (2003).
18. Rishton, G. M. Reactive compounds and *in vitro* false positives in HTS. *Drug Discov. Today* **2**, 382–384 (1997).
19. Rishton, G. M. Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discov. Today* **8**, 86–96 (2003).
20. Swinney, D. C. Biochemical mechanism of drug action: what does it take for success? *Nature Rev. Drug Discov.* **3**, 801–808 (2004).
21. Baurin, N. *et al.* Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds. *J. Chem. Inf. Comput. Sci.* **44**, 643–651 (2004).
22. Andrews, P. R., Craik, D. J. & Martin, J. L. Functional group contributions to drug-receptor interactions. *J. Med. Chem.* **27**, 1648–1657 (1984).
23. Hann, M. M., Leach, A. R. & Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* **41**, 856–864 (2001).
24. Williams, M. A return to the fundamentals of drug discovery. *Curr. Opin. Investig. Drugs* **5**, 29–33 (2004).
25. Horrobin, D. F. Modern biomedical research: an internally self-consistent universe with little contact with medical reality. *Nature Rev. Drug Discov.* **2**, 151–154 (2003).
26. Morphy, R., Kay, C. & Rankovic, Z. From magic bullets to designed multiple ligands. *Drug Discov. Today* **9**, 641–651 (2004).
27. Lipper, R. A. How can we optimize selection of drug development candidates from many compounds at the discovery stage? *Mod. Drug Discov.* **2**, 55–60 (1999).
28. Van Gestel, S. & Schuermans, V. Thirty-three years of drug discovery and research with Dr. Paul Janssen. *Drug Dev. Res.* **8**, 1–13 (1986).
29. Raju, T. N. The Nobel chronicles. *Lancet* **355**, 1022 (2000).
30. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nature Rev. Drug Discov.* **1**, 727–730 (2002).
31. Hopkins, A. L. & Groom, C. R. Target analysis: a priori assessment of druggability. *Ernst Schering Res. Found. Workshop* **42**, 11–17 (2003).
32. Arkin, M. R. & Wells, J. A. Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. *Nature Rev. Drug Discov.* **3**, 301–317 (2004).
33. Teague, S. J. Implications of protein flexibility for drug discovery. *Nature Rev. Drug Discov.* **2**, 527–541 (2003).
34. Rader, C. Antibody libraries in drug and target discovery. *Drug Discov. Today* **6**, 36–43 (2001).
35. Graddis, T. J. *et al.* Designing proteins that work using recombinant technologies. *Curr. Pharm. Biotechnol.* **3**, 285–297 (2002).
36. Zambrowicz, B. P. & Sands, A. T. Modeling drug action in the mouse with knockouts and RNA interference. *Drug Discov. Today: TARGETS* **3**, 198–207.
37. Shokat, K. & Velleca, M. Novel chemical genetic approaches to the discovery of signal transduction inhibitors. *Drug Discov. Today* **7**, 872–879 (2002).
38. Hannon, G. J. & Rossi, J. J. Unlocking the potential of the human genome with RNA interference. *Nature* **431**, 371–378 (2004).
39. Lobato, M. N. & Rabbitts, T. H. Intracellular antibodies as specific reagents for functional ablation: future therapeutic molecules. *Curr. Mol. Med.* **4**, 519–528 (2004).
40. Rees, D. C., Congreve, M., Murray, C. W. & Carr, R. Fragment-based lead discovery. *Nature Rev. Drug Discov.* **3**, 660–672 (2004).
41. Shuker, S. B., Hajduk, P. J., Meadows, R. P. & Fesik, S. W. Discovering high-affinity ligands for proteins: SAR by NMR. *Science* **274**, 1531–1534 (1996).
42. Swayze, E. E. *et al.* SAR by MS: a ligand based technique for drug lead discovery against structured RNA targets. *J. Med. Chem.* **45**, 3816–3819 (2002).
43. Nienaber, V. L. *et al.* Discovering novel ligands for macromolecules using X-ray crystallographic screening. *Nature Biotechnol.* **18**, 1105–1108 (2000).
44. Lesuisse, D. *et al.* SAR and X-ray. A new approach combining fragment-based screening and rational drug design: application to the discovery of nanomolar inhibitors of Src SH2. *J. Med. Chem.* **45**, 2379–2387 (2002).
45. Blundell, T. L., Jhoti, H. & Abell, C. High-throughput crystallography for lead discovery in drug design. *Nature Rev. Drug Discov.* **1**, 45–54 (2002).
46. Kolb, H. C. & Sharpless, K. B. The growing impact of click chemistry on drug discovery. *Drug Discov. Today* **8**, 1128–1137 (2003).
47. Kolb, H. C., Finn, M. G. & Sharpless, K. B. Click chemistry: diverse chemical function from a few good reactions. *Angew. Chem. Int. Edn Engl.* **40**(11), 2004–2021 (2001).
48. Erlanson, D. A. *et al.* Site-directed ligand discovery. *Proc. Natl Acad. Sci. USA* **97**, 9367–9372 (2000).
49. Erlanson, D. A., Wells, J. A. & Braisted, A. C. Tethering: fragment-based drug discovery. *Annu. Rev. Biophys. Biomol. Struct.* **33**, 199–223 (2004).
50. Weininger, D. in *Encyclopedia of Computational Chemistry* (eds Von Ragué Schleyer, P. *et al.*) 425–530 (Wiley, New York, 1998).
51. Oprea, T. I. & Gottfries, J. Chemography: the art of navigating in chemical space. *J. Comb. Chem.* **3**, 157–166 (2001).
52. Oprea, T. I. Chemical space navigation in lead discovery. *Curr. Opin. Chem. Biol.* **6**, 384–389 (2002).
53. Vieth, M. *et al.* Characteristic physical properties and structural fragments of marketed oral drugs. *J. Med. Chem.* **47**, 224–232 (2004).
54. Ajay, A., Walters, W. P. & Murcko, M. A. Can we learn to distinguish between ‘drug-like’ and ‘nondrug-like’ molecules? *J. Med. Chem.* **41**, 3314–3324 (1998).
55. Wang, J. & Ramnarayan, K. Towards designing drug-like libraries: a novel computational approach for prediction of drug feasibility of compounds. *J. Comb. Chem.* **1**, 524–533 (1999).
56. Walters, W. P., Ajay & Murcko, M. A. Recognizing molecules with drug-like properties. *Curr. Opin. Chem. Biol.* **3**, 384–387 (1999).
57. Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **44**, 3–25 (2000).
58. Podlogar, B. L., Muegge, I. & Brice, L. J. Computational methods to estimate drug development parameters. *Curr. Opin. Drug Discov. Devel.* **4**, 102–109 (2001).
59. Muegge, I., Heald, S. L. & Brittelli, D. Simple selection criteria for drug-like chemical matter. *J. Med. Chem.* **44**, 1841–1846 (2001).
60. Veber, D. F. *et al.* Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **45**, 2615–2623 (2002).
61. Proudfoot, J. R. Drugs, leads, and drug-likeness: an analysis of some recently launched drugs. *Bioorg. Med. Chem. Lett.* **12**, 1647–1650 (2002).
62. Walters, W. P. & Murcko, M. A. Prediction of ‘drug-likeness’. *Adv. Drug Deliv. Rev.* **54**, 255–271 (2002).
63. Egan, W. J., Walters, W. P. & Murcko, M. A. Guiding molecules towards drug-likeness. *Curr. Opin. Drug Discov. Devel.* **5**, 540–549 (2002).
64. Muegge, I. Selection criteria for drug-like compounds. *Med. Res. Rev.* **23**, 302–321 (2003).
65. Lajiness, M. S., Vieth, M. & Erickson, J. Molecular properties that influence oral drug-like behavior. *Curr. Opin. Drug Discov. Devel.* **7**, 470–477 (2004).
66. Horn, J. R. & Shoichet, B. K. Allosteric inhibition through core disruption. *J. Mol. Biol.* **336**, 1283–1291 (2004).

Acknowledgements We thank R. W. Spencer, J. Everett and J. Mason for discussions and advice during the preparation of this manuscript.

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Virtual screening of chemical libraries

Brian K. Shoichet

Department of Pharmaceutical Chemistry, University of California, 600 16th Street, San Francisco, California 94143-2240, USA
(e-mail: shoichet@cgl.ucsf.edu)

Virtual screening uses computer-based methods to discover new ligands on the basis of biological structures. Although widely heralded in the 1970s and 1980s, the technique has since struggled to meet its initial promise, and drug discovery remains dominated by empirical screening. Recent successes in predicting new ligands and their receptor-bound structures, and better rates of ligand discovery compared to empirical screening, have re-ignited interest in virtual screening, which is now widely used in drug discovery, albeit on a more limited scale than empirical screening.

The dominant technique for the identification of new lead compounds in drug discovery is the physical screening of large libraries of chemicals against a biological target (high-throughput screening). An alternative approach, known as virtual screening, is to computationally screen large libraries of chemicals for compounds that complement targets of known structure, and experimentally test those that are predicted to bind well. Such receptor-based virtual screening faces several fundamental challenges, including sampling the various conformations of flexible molecules and calculating absolute binding energies in an aqueous environment. Nevertheless, the field has recently had important successes: new ligands have been predicted along with their receptor-bound structures — in several cases with hit rates (ligands discovered per molecules tested) significantly greater than with high-throughput screening. Even with its current limitations, virtual screening accesses a large number of possible new ligands, most of which may then be simply purchased and tested. For those who can tolerate its false-positive and false-negative predictions, virtual screening offers a practical route to discovering new reagents and leads for pharmaceutical research.

Problems with virtual screening

A founding idea in molecular biology was that biological function follows from molecular form. If you knew the molecular structure of a receptor — defined here as a biological macromolecule that converts ligand binding into an activity — you could understand and predict its function. This notion has underpinned a 70-year project to determine receptor structures to atomic resolution. From the early X-ray diffraction studies of pepsin and of haemoglobin, to those of macromolecular assemblies like the ribosome and to structural genomics, the taxonomic part of this enterprise (that is, cataloguing receptor structures) has been extraordinarily successful. But still largely unfulfilled is the promise of exploiting receptor structures to discover new ligands that modulate the activities of these molecules and macromolecular assemblies.

As early as the mid-1970s, investigators suggested that computational simulations of receptor structures and the chemical forces that govern their interactions would enable 'structure-based' ligand design and discovery^{1,2}. Ligands could be designed on the basis of the receptor structure alone, which would free medicinal chemistry from the tyranny of empirical screening, substrate-based design and incremental modification. Since then, structure-based design has contributed to and even motivated the development of marketed drugs^{3,4}, such as the human immunodeficiency virus (HIV) protease inhibitor Viracept and the

anti-influenza drug Relenza, typically through cycles of modification and subsequent experimental structure determination. Computational modelling has been used extensively in these efforts^{5,6} and indeed in non-receptor-based methods; for example, when searching for new ligands on the basis of their chemical similarity to a known ligand or when matching candidate molecules to a 'pharmacophore' that represents the chemical properties of a series of known ligands⁷. But until recently there have been few instances of completely new ligands (not resembling those previously known) discovered directly from receptor-based computation. Although there are now many more and much better receptor structures than there were in the 1970s and 1980s, and computer speed has grown exponentially, drug discovery and chemical biology remain dominated by empirical screening and substrate-based design.

Three problems have impeded progress in receptor-guided explorations of ligand chemistry. First, chemical space is vast but most of it is biologically uninteresting: blank, lightless galaxies exist within it into which good ideas at their peril wander. Constraining the number of chemical compounds that are searched to biologically relevant and synthetically accessible molecules remains an area of active research. Second, receptor structures are complicated, resembling "tangled knot(s) of viscera"⁸. They consist of several thousand atoms, each of which is more or less free to move, and they frequently change shape and solvent structure upon binding to a ligand. To predict what molecules might be recognized by a given receptor, energetically accessible receptor and ligand conformations should be calculated. Unfortunately, the number of possible conformations rises exponentially with the number of rotatable bonds, of which there are thousands in a protein–ligand complex, and the full sampling of conformations involves a set of computational problems for which no general solution is known. Third, calculating ligand–receptor binding energies is difficult⁹. Binding affinity in an aqueous environment is determined by the solvation energies of the individual molecules (high solvation energies typically disfavour binding), and by the interaction energies between them (high interaction energies favour binding). Solvation and interaction energies are both typically much larger in magnitude than the net affinity, making calculation of the latter problematic. Although it has been possible to calculate accurately the differential affinity between two related ligands using thermodynamic integration methods, doing so is time consuming. Calculating the absolute affinities for many thousands of unrelated molecules necessary to encode new chemical functionality remains beyond our reach. So in principle, it could be argued that structure-based computational screens for new ligands do not work at all.

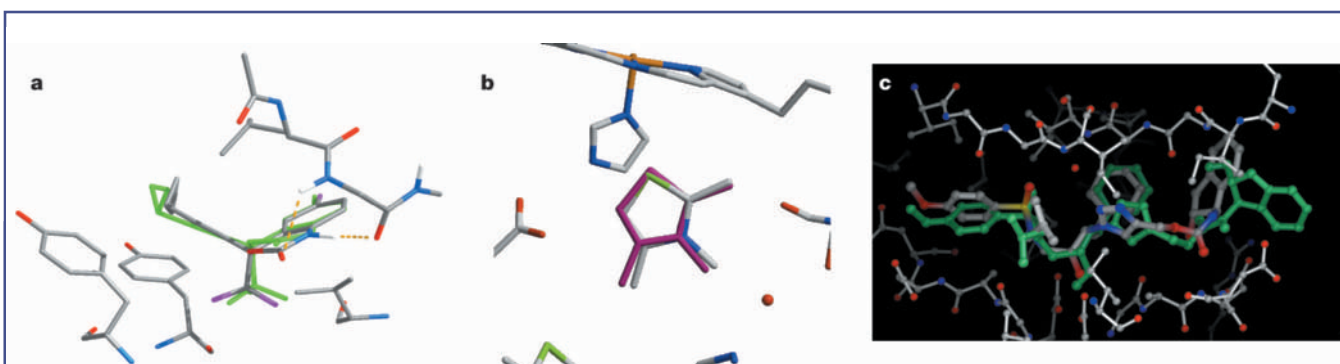


Figure 1 Complexes predicted from virtual screening compared to X-ray crystallographic structures that were subsequently determined. **a**, Predicted (carbons in grey) and experimental (green) structures for Sustiva in HIV reverse transcriptase¹⁰. **b**, Predicted (magenta) and experimental (carbons in grey) structures of 2,3,4-

trimethylthiazole in the W191G cavity of cytochrome *c* peroxidase¹¹. **c**, Predicted (green)¹² and experimental structure (carbons in grey) of an amprevir mimic in HIV protease (ligands with thick bonds, enzyme residues with thin bonds; structure determined by A. Wlodawer, A. Olson, personal communication).

Successes from virtual screening

However, genuinely novel ligands have been discovered using structure-based computation. Recently, the structures of known ligands in complex with their receptors have been correctly predicted computationally using the structures of the independent receptor and ligand molecules^{10–12} (Fig. 1). From the standpoint of exploring chemical space, computational screens of chemical databases have identified new ligands for over 50 receptors of known or even, in some cases, computer-modelled structures^{13,14} (for reviews of recent studies and methods see refs 15 and 16). In these virtual or ‘docking’ screens, large libraries of organic molecules are docked into receptor structures and ranked by the calculated affinity (Fig. 2). Although the energy calculations are crude, the compounds in the library are readily available, making experimental testing easy and false-positives tolerable⁵.

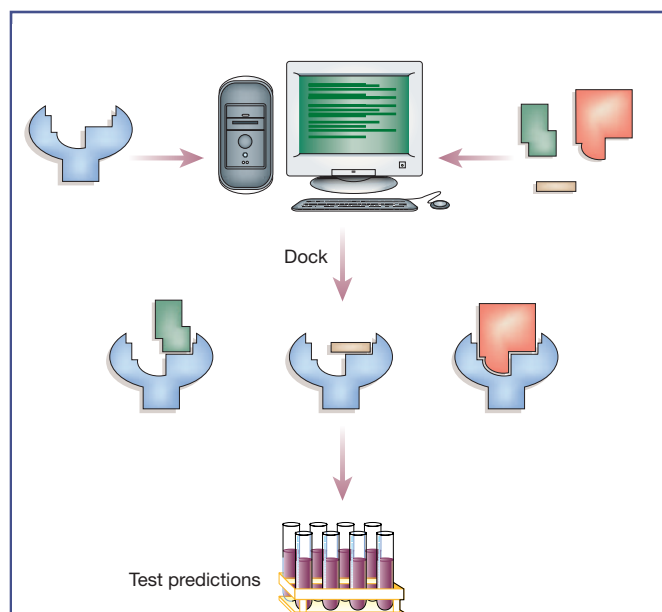


Figure 2 Virtual screening for new ligands. Large libraries of available, often purchasable, compounds are docked into the structure of receptor targets by a docking computer program. Each compound is sampled in thousands to millions of possible configurations and scored on the basis of its complementarity to the receptor. Of the hundreds of thousands of molecules in the library, tens of top-scoring predicted ligands (hits) are subsequently tested for activity in an experimental assay.

Even relatively simple receptor-based constraints can improve the likelihood of finding ligands from among the many possible structures in a library, if only by screening out those that are unlikely to bind the receptor¹⁷. In library design, for instance, pre-calculation of possible side chains that would complement a receptor structure resulted in structure-based libraries that were tenfold more likely to contain ligands than random¹⁸ or diverse¹⁷ libraries constructed at the same time. Similarly, virtual and high-throughput screening have been deployed simultaneously to discover new ligands from libraries of several-hundred-thousand diverse molecules. The virtual screens had ‘hit rates’ (defined as the number of compounds that bind at a particular concentration divided by the number of compounds experimentally tested) that were 100-fold to 1,000-fold higher than those achieved by empirical screens^{19,20} (Table 1); intriguingly, each technique discovered classes of ligands that the other technique had overlooked¹⁹, suggesting that the two screening approaches (virtual and empirical) can be complementary.

In a few cases the structures of the new ligands in complex with the receptors have been subsequently determined experimentally — typically by X-ray crystallography. Although the docking-derived hits are very different from natural ligands for a given receptor, they often bind at the active site, interacting with conserved receptor groups, as predicted by the docking program^{21–24} (Fig. 3). From a molecular recognition perspective, this suggests that the structural ‘code’ for binding is plastic in that multiple ligand scaffolds can be recognized by the same receptor site. Methodologically, these structures suggest that although virtual screens are plagued by false-positives, in favourable circumstances they can predict genuinely novel ligands and do so for the right reasons.

How can these successes be reconciled with the field’s methodological weaknesses? Virtual screening avoids the problem of broad searches of chemical space by restricting itself to libraries of specific, accessible compounds (often those that can simply be purchased). This avoids costly syntheses and restricts the search to compounds that are interesting enough biologically to have been previously made, albeit for another reason. Filters may be applied to ensure that the library meets some standard of biological relevance or ‘drug-likeness’^{25,26}. Progress in both the number and quality of molecules in docking libraries has contributed to the increasingly drug-like character of docking hits in recent studies¹⁹. Although the problems of sampling molecular conformations and of calculating affinities remain acute, progress has been made both algorithmically¹⁶ and in the computer resources available for these calculations. Moreover, we can define success in virtual screening as ‘finding some interesting new ligands’, and not as ‘correctly ranking all the molecules in the library’ or ‘finding all the possible ligands in a library’. Virtual

Table 1 Hit rates and drug-like properties for inhibitors discovered with high-throughput and virtual screening against the enzyme PTP-1B (ref. 19)

Technique	Compounds tested	Hits with $IC_{50} < 100 \mu\text{M}$	Hits with $IC_{50} < 10 \mu\text{M}$	Lipinski compliant hits*	Hit rate†
HTS	400,000	85	6	23	0.021%
Docking	365‡	127	18	57	34.8%

*Number of 100 μM or better inhibitors that passed all four of the drug-like criteria identified in Lipinski's 'rule of five'²⁵; †The number of compounds experimentally tested divided by the number of compounds with IC_{50} values of 100 μM or less; ‡The number of top-scoring docking hits that were experimentally tested; IC_{50} , The concentration of inhibitor at which the enzyme is 50% inhibited.

screening thus adopts the same logic as high-throughput screening: as long as some interesting ligands are found, false-negatives are tolerated. Indeed, the two techniques, because of their emphasis on large libraries, share other similarities: both accept limited accuracy in return for screening on a large scale; both look to enrich a list of likely-but-not-certain candidates for further quantitative study; and both are dogged by curious false-positive hits²⁷. Although high-throughput screening remains the dominant technique, virtual screening is now commonly used in pharmaceutical research.

Finally, it must be admitted that these successes retain an episodic character. Even expert practitioners are frequently surprised and sometimes disappointed. Geometries of true ligands may be slightly (Fig. 3e)²⁸ or conspicuously (Fig. 3f)²⁹ mis-predicted and hit rates can vary greatly. We have had hit rates as high as 35% (ref. 19) against an enzyme, protein tyrosine phosphatase 1B (PTP1B), with which we

had little experience, and as low as 5% (ref. 22) against an enzyme, AmpC β -lactamase, that we had studied intensely. For many medicinal chemists and structural biologists, such unpredictability lends a whiff of sulphur to an enterprise that has been advertised as 'rational drug design'.

Prospects

Notwithstanding these caveats, virtual screening will be an ever-more important tool for exploring biologically relevant chemical space. Large high-throughput screens have liabilities of their own, and are inaccessible to many investigators (although this will begin to change with the advent of screening resource centres³⁰). In contrast, virtual screening processes large libraries (in principle, libraries that are larger than any library used by empirical screening) and any receptor for which there is a structure at little cost. What advances

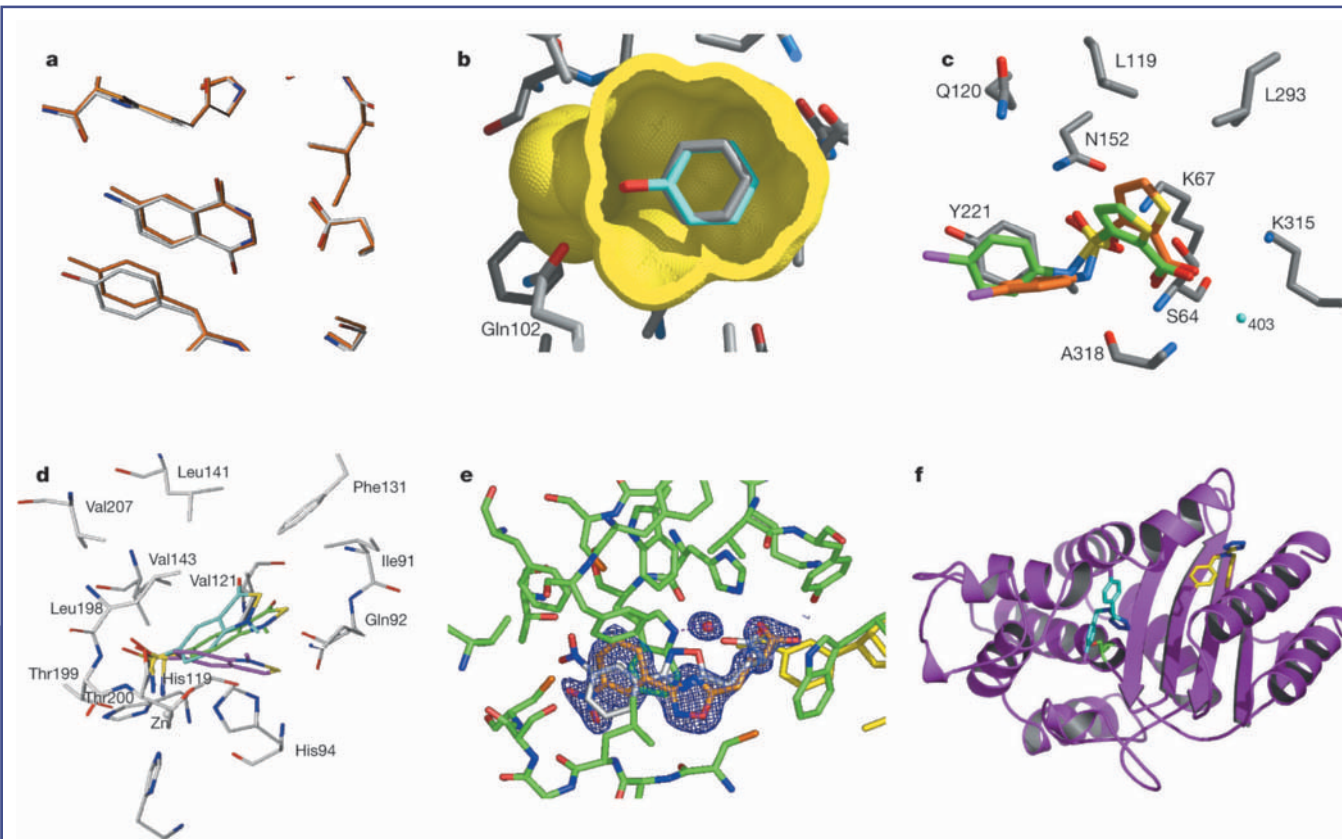


Figure 3 Comparing the structures of new ligands predicted from virtual screening to the structures subsequently determined experimentally. **a**, The docked (carbons in orange) versus the crystallographic structure (carbons in grey) of the 8.3 μM inhibitor 4-aminophthalhydrazide bound to transfer RNA guanine transglycosylase (ligand in the centre surrounded by enzyme residues)²¹. **b**, The docked (carbons in cyan) versus the crystallographic structure (carbons in grey) of the 100 μM ligand phenol bound to a cavity site in T4 lysozyme (ligand in the centre surrounded by the molecular surface of the surrounding protein residues)²⁴. **c**, The docked (carbons in green) versus the crystallographic structure (carbons in red) of the 26 μM inhibitor 3-((4-chloroanilino)-sulphonyl)-thiophene-2-carboxylate bound to AmpC β -lactamase (enzyme carbons in grey)²². **d**, The docked (carbons in magenta), re-scored (carbons

in cyan) and crystallographic (carbons in grey) structures of a 0.25 μM inhibitor bound to carbonic anhydrase (enzyme carbons in grey)²³. Oxygen atoms in red, sulphurs in yellow, nitrogens in blue. **e**, The docked (ligand carbons in orange) versus the crystallographic structure (ligand carbons in grey) for a new inhibitor of aldose reductase (enzyme carbons in green). Electron density maps for the ligand are shown in blue. The ordered water (red sphere) observed in the experimental structure was not considered in the docking²⁸ (H. Steuber and G. Klebe, unpublished work). **f**, The docked (carbons in cyan) versus the crystallographic structure (carbons in yellow) of the new inhibitor of TEM-1 β -lactamase (enzyme in magenta)²⁹. The experimentally observed binding mode — 16 Å from the active site targeted in the docking calculations — occurs in a cryptic site absent from the native structure.

might be anticipated to make virtual screening reliable and accessible enough to be widely used?

Improved sampling and 'scoring functions' (calculations of ligand–receptor energetics) will undoubtedly help. The good news is that the fundamentals of molecular interactions are well understood, and so the field has a clear way forward. But the challenge, as always, will be to implement good physical models for hundreds of thousands of possible ligands, each one sampled in many thousands of possible receptor complexes. Indeed, accurate calculation of absolute binding affinity in screens of large, diverse libraries will remain beyond us for the foreseeable future; even predicting the rank order of affinity for disparate ligands in a hit list will be difficult. What we may anticipate are improved explorations of conformational states for ligand and receptor, and scoring functions that use more sophisticated models of solvation and a better balance of electrostatic and non-polar terms. An interesting strategy will be the use of higher-level, typically much slower methods to re-score initial hits from virtual screening, using the screening calculation as a fast first filter³¹. From these we can hope for better hit rates and better predictions of geometries²³ (Fig. 3d), which are the first and most important goals of virtual screening.

To bring virtual screening to a wide community it will be important to democratize the resources on which it depends. Receptor structures are already available through the Protein Data Bank or PDB (for experimental structures), and through databases such as MODBASE (for a much larger number of structures from computer-based modelling³²). Several groups provide docking programs without charge to the academic community, although these programs often require some effort to learn. Programs less demanding of expert knowledge, perhaps as a web-accessible resource, would bring docking to many interested non-specialists. Finally, community-accessible chemical libraries are needed. The National Cancer Institute (NCI) provides calculated structures for about 140,000 of its compounds, and will provide at least some of these for experimental testing (<http://cactus.nci.nih.gov/>). MDL Inc. sells the Available Chemicals Directory (ACD; http://www.mdll.com/products/experiment/available_chem_dir/index.jsp) of commercially available compounds and the ACD-SC for screening collections. To use these libraries in docking screens, molecular properties such as protonation, charge, stereochemistry, accessible conformations and solvation must be calculated. Even details such as stereochemistry, tautomerization and protonation, which we frequently take for granted, are often ambiguous, or can change on binding to a receptor. Recently, about one million commercially accessible molecules have become available through the ZINC database (<http://blaster.docking.org/zinc/>). ZINC is a free, web-accessible database constructed with docking, sub-structure searching and compound purchasing in mind.

In the immediate future, virtual screening is mature enough to benefit from an aggressive programme of experimental testing. As more docking predictions are evaluated, and sometimes falsified, the methods will improve, especially if care is taken to remove the false-positives that have plagued both high-throughput and virtual screening²⁷. Subsequent solution of receptor–ligand complex structures will be particularly informative; so far, too few of these have been determined. For those who can tolerate its false-positives, structure-based virtual screening is reliable enough to justify its use in active ligand discovery projects, providing an important complementary approach to empirical screening. For some projects, especially those centred in academic laboratories, virtual screening will be the best way to access a large chemical space without the

commitment in time, material and infrastructure that an empirical screen demands. □

doi:10.1038/nature03197

- Beddell, C. R., Goodford, P. J., Norrington, F. E., Wilkinson, S. & Wootton, R. Compounds designed to fit a site of known structure in human haemoglobin. *Br. J. Pharmacol.* **57**, 201–209 (1976).
- Cohen, S. S. A strategy for the chemotherapy of infectious disease. *Science* **197**, 431–432 (1977).
- Itzstein, M. V. *et al.* Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* **363**, 418–423 (1993).
- Varney, M. D. *et al.* Crystal-structure-based design and synthesis of Benz[cd]indole-containing inhibitors of thymidylate synthase. *J. Med. Chem.* **35**, 663–676 (1992).
- Kuntz, I. D. Structure-based strategies for drug design and discovery. *Science* **257**, 1078–1082 (1992).
- Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **303**, 1813–1818 (2004).
- Stahura, F. L. & Bajorath, J. Virtual screening methods that complement HTS. *Comb. Chem. High Throughput Screen* **7**, 259–269 (2004).
- Perutz, M. F. The hemoglobin molecule. *Sci. Am.* **211**, 64–76 (1964).
- van Gunsteren, W. F. & Berendsen, H. J. C. Computer simulation of molecular dynamics: methodology, applications, and perspectives in chemistry. *Angew. Chem. Int. Ed. Engl.* **29**, 992–1023 (1990).
- Rizzo, R., Wang, D., Tirado-Rives, J. & Jorgensen, W. Validation of a model for the complex of HIV-1 reverse transcriptase with sustiva through computation of resistance profiles. *J. Am. Chem. Soc.* **122**, 12898–12900 (2000).
- Rosenfeld, R. J. *et al.* Automated docking of ligands to an artificial active site: augmenting crystallographic analysis with computer modeling. *J. Comput. Aided Mol. Des.* **17**, 525–536 (2003).
- Brik, A. *et al.* Rapid diversity-oriented synthesis in microtiter plates for *in situ* screening of HIV protease inhibitors. *Chembiochem.* **4**, 1246–1248 (2003).
- Schapiro, M. *et al.* Discovery of diverse thyroid hormone receptor antagonists by high-throughput docking. *Proc. Natl Acad. Sci. USA* **100**, 7354–7359 (2003).
- Evers, A. & Klebe, G. Ligand-supported homology modeling of G-protein-coupled receptor sites: models sufficient for successful virtual screening. *Angew. Chem. Int. Ed. Engl.* **43**, 248–251 (2004).
- Shoichet, B. K., McGovern, S. L., Wei, B. & Irwin, J. J. Lead discovery using molecular docking. *Curr. Opin. Chem. Biol.* **6**, 439–446 (2002).
- Schneidman-Duhovny, D., Nussinov, R. & Wolfson, H. J. Predicting molecular interactions *in silico*. II. Protein–protein and protein–drug docking. *Curr. Med. Chem.* **11**, 91–107 (2004).
- Wyss, P. C. *et al.* Novel dihydrofolate reductase inhibitors. Structure-based versus diversity-based library design and high-throughput synthesis and screening. *J. Med. Chem.* **46**, 2304–2312 (2003).
- Kick, E. K. *et al.* Structure-based design and combinatorial chemistry yield low nanomolar inhibitors of cathepsin D. *Chem. Biol.* **4**, 297–307 (1997).
- Doman, T. N. *et al.* Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* **45**, 2213–2221 (2002).
- Paiva, A. M. *et al.* Inhibitors of dihydrodipicolinate reductase, a key enzyme of the diaminopimelate pathway of *Mycobacterium tuberculosis*. *Biochim. Biophys. Acta* **1545**, 67–77 (2001).
- Grädler, U. *et al.* A new target for shigellosis: rational design and crystallographic studies of inhibitors of tRNA-guanine transglycosylase. *J. Mol. Biol.* **306**, 455–467 (2001).
- Powers, R. A., Morandi, F. & Shoichet, B. K. Structure-based discovery of a novel, noncovalent inhibitor of AmpC beta-lactamase. *Structure (Camb.)* **10**, 1013–1023 (2002).
- Gruneberg, S., Stubbs, M. T. & Klebe, G. Successful virtual screening for novel inhibitors of human carbonic anhydrase: strategy and experimental confirmation. *J. Med. Chem.* **45**, 3588–3602 (2002).
- Wei, B. Q., Baase, W. A., Weaver, L. H., Matthews, B. W. & Shoichet, B. K. A model binding site for testing scoring functions in molecular docking. *J. Mol. Biol.* **322**, 339–355 (2002).
- Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **23**, 3–25 (1997).
- Oprea, T. I. Current trends in lead discovery: are we looking for the appropriate properties? *Mol. Divers.* **5**, 199–208 (2002).
- McGovern, S. L., Caselli, E., Grigorieff, N. & Shoichet, B. K. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem.* **45**, 1712–1722 (2002).
- Krämer, O., Hazemann, L., Podjarny, A. D. & Klebe, G. Virtual screening for inhibitors of human aldose reductase. *Proteins* **55**, 814–823 (2004).
- Horn, J. R. & Shoichet, B. K. Allosteric inhibition through core disruption. *J. Mol. Biol.* **336**, 1283–1291 (2004).
- Kaiser, J. NIH Gears up for chemical genomics. *Science* **304**, 1728 (2004).
- Kalyanaraman, C., Bernacki, K. & Jacobson, M. P. Virtual screening against highly charged active sites: Identifying substrates of alpha-beta barrel enzymes. *Biochemistry* in the press.
- Pieper, U., Eswar, N., Stuart, A. C., Ilyin, V. A. & Salí, A. MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res.* **30**, 255–259 (2002).

Acknowledgements I thank G. Klebe, A. Olson, and W. Jorgensen for contributing figures and comments, and I. D. Kuntz, M. Jacobson, A. Salí, K. Dill and J. Irwin for many insightful conversations. My laboratory's research in docking is supported by NIGMS.

Competing interests statement The author declares competing financial interests: details accompany the paper on www.nature.com/nature.

Tomographic imaging of molecular orbitals

J. Itatani^{1,2}, J. Levesque^{1,3}, D. Zeidler¹, Hiromichi Niikura^{1,4}, H. Pépin³, J. C. Kieffer³, P. B. Corkum¹ & D. M. Villeneuve¹

¹National Research Council of Canada, 100 Sussex Drive, Ottawa, Ontario K1A 0R6, Canada

²University of Ottawa, 150 Louis Pasteur, Ottawa, Ontario K1N 6N5, Canada

³INRS-Énergie et Matériaux, 1650 boulevard Lionel-Boulet, CP 1020, Varennes, Québec J3X 1S2, Canada

⁴PRESTO, Japan Science and Technology Agency, 4-1-8 Honcho Kawaguchi Saitama, 332-0012, Japan

Single-electron wavefunctions, or orbitals, are the mathematical constructs used to describe the multi-electron wavefunction of molecules. Because the highest-lying orbitals are responsible for chemical properties, they are of particular interest. To observe these orbitals change as bonds are formed and broken is to observe the essence of chemistry. Yet single orbitals are difficult to observe experimentally, and until now, this has been impossible on the timescale of chemical reactions. Here we demonstrate that the full three-dimensional structure of a single orbital can be imaged by a seemingly unlikely technique, using high harmonics generated from intense femtosecond laser pulses focused on aligned molecules. Applying this approach to a series of molecular alignments, we accomplish a tomographic reconstruction of the highest occupied molecular orbital of N₂. The method also allows us to follow the attosecond dynamics of an electron wave packet.

The electrons that make up molecules are organized by energy in orbitals^{1,2}. Although total electron density in molecules is routinely measured by X-ray diffraction or electron scattering, only two methods are able to ‘see’ the highest occupied molecular orbitals (HOMOs)—electron momentum spectroscopy and scanning tunnelling microscopy. Electron momentum spectroscopy³ is an (*e*, 2*e*) scattering technique that can determine the radially averaged density of the outermost valence electrons. Scanning tunnelling microscopy⁴ gives the electron density, distorted by surface states. We show that high harmonic emission from molecules^{5–11} allows the three-dimensional shape of the highest electronic orbital to be measured, including the relative phase of components of the wavefunction. Our results may be compared with predictions of the various models that are used to describe many-electron systems¹², such as the Hartree–Fock, the Kohn–Sham and the Dyson approaches.

If we can measure orbitals with femtosecond laser technology, we can also observe orbital changes occurring on the timescale of chemical reactions. The next challenge will be to measure attosecond bound-state electron dynamics. We show that imaging orbitals can be extended to imaging coherent electron motion in atoms or molecules. Quantum mechanically, electron motion is described by a coherent superposition of electronic states. This is the closest that quantum mechanics allows to imaging the ‘Bohr orbital motion’ of electrons.

Tomographic imaging of a molecular orbital is achieved in three crucial steps. (1) Alignment of the molecular axis in the laboratory frame. (2) Selective ionization of the orbital. (3) Projection of the state onto a coherent set of plane waves.

Alignment of gas-phase molecules

Computed tomography refers to the technology of retrieving sectional images of an object, such as a human body, from a series of one-dimensional projections¹³. For a projection direction fixed in the laboratory frame, the object must be rotated. Thus, we begin by discussing how molecules can be aligned¹⁴.

A relatively intense, non-resonant laser pulse induces a Stark shift of the ground state of molecules that depends on the angle of the molecular axis to the laser polarization¹⁵. It has been demonstrated that gas-phase molecules can be aligned adiabatically¹⁶, aligned in three dimensions¹⁷ or oriented¹⁸. We use non-adiabatic

alignment^{19,20}. It results in field-free alignment of linear molecules well after the aligning pulse has terminated. A 60-fs laser pulse produces a rotational wave packet in N₂. This wave packet periodically rephases, giving periods when the molecular axes are aligned in space. The probe pulse coincides with the first half-revival at 4.1 ps (ref. 20). The degree of angular alignment is good enough to see a difference in the harmonic spectra for a 5° rotation of the alignment direction.

High harmonic generation from molecules

High harmonics are produced by ionizing atoms or molecules. We choose to operate in the tunnel ionization regime²¹ because the rate of tunnel ionization depends exponentially on the ionization potential. Hence the orbital with the lowest ionization potential (the HOMO in an unexcited molecule) is preferentially selected. This is the second requirement for tomographic imaging of a single molecular orbital.

The third requirement for tomography is that we can take a series of projected images of the molecular orbital. For this we record high harmonic emission generated from the molecule itself. High harmonics have been extensively studied from atoms, whereas harmonics from molecules have been studied with isotropic distributions^{5,6} or with weak alignment⁷. It has been observed^{8–11} both experimentally and numerically, that harmonic emission from aligned molecules is sensitive to both the alignment angle and to the spatial structure of the electronic wavefunction. These observations provide the basis for our analysis.

High harmonic generation is understood using the three-step quasi-static model²²—(1) tunnel ionization by an intense low-frequency laser field, (2) acceleration of the free electron, and (3) re-collision. Tunnel ionization coherently transfers a part of the bound-state electron wavefunction to the continuum. Once free, the strong laser field dominates the motion of the continuum electron wave packet. First it propagates away from the molecule (Fig. 1a) and is then driven back after the field reverses its direction. Figure 1b shows the re-colliding wave packet expanding laterally. By the time of re-collision, the electron wave packet has a typical transverse 1/*e* width of 9 Å for 800-nm laser fields^{23,24}. Because the wave packet is much larger than the size of small molecules (typically ~1 Å, shown by the lines in Fig. 1b), the molecule sees essentially a plane electron wave re-colliding. We will see that the planar nature of the re-collision electron allows us to experimentally retrieve a

one-dimensional projection of the ground-state wavefunction.

The laser field shears the initial continuum wave packet in such a way that the low kinetic energy (long de Broglie wavelength) part of the wave packet re-collides first. Over the next half laser period, it chirps to high kinetic energy (short wavelength) and then back to low energy²⁵. In our experiment, the highest-energy electron has a wavelength of ~ 1.5 Å. This wavelength is related to the high harmonic cut-off. The broad range of electron wavelengths allows us to experimentally retrieve the spatial shape of the bound-state wavefunction.

Figure 2 illustrates the final step in the harmonic generation process. Figure 2a shows a bound-state wavefunction, Ψ_g . A fraction of the electron wavefunction tunnels from this orbital. The re-colliding electron wave packet, Ψ_c , shown as a plane wave in Fig. 2b, overlaps the remaining portion of the initial wavefunction. The coherent addition of the two wavefunctions induces a dipole, seen as the asymmetric localization of electron density in Fig. 2c. The induced dipole oscillates as the continuum wavefunction propagates. It is this oscillating dipole that emits high harmonic radiation. The instantaneous frequency of the oscillating dipole is related to the electron kinetic energy by $\hbar\omega = E_k$ (refs 22, 25). In other words, the electron wave packet and the emitted photons are

related by energy conservation, and are mutually coherent. We do not use the more usual expression²², $\hbar\omega = E_k + I_p$, where I_p is the ionization potential and E_k is the instantaneous kinetic energy of the electron in the continuum at the time of re-collision calculated for a delta function potential, because here we are concerned with the electron kinetic energy seen by the bound-state electron wavefunction.

Formally, we are measuring the transition dipole moment from the highest occupied state to a set of continuum wavefunctions. Mathematically, the harmonic spectrum from a single atom or molecule is given by the Fourier transform of the dipole acceleration²⁶. We relate the dipole acceleration spectrum to the dipole moment using the slowly varying envelope approximation (see Methods section). The spectral intensity (I) and phase of harmonics (ϕ) at frequency ω can respectively be written as $I(\omega) \propto \omega^4 |d(\omega)|^2$ and $\phi(\omega) = \arg[d(\omega)]$, where the transition dipole moment in the spectral domain is $d(\omega) = a[k(\omega)] \int \Psi_g(\mathbf{r})(\mathbf{er}) \exp[ik(\omega)x] d\mathbf{r}$. Here the electron wave packet seen by the molecule is expanded in a superposition of plane waves as $\Psi_c = \int a(k) \exp[ik(\omega)x] dk$, $k(\omega)$ is the wavenumber (momentum) of the electron corresponding to harmonic frequency ω , and $a[k(\omega)]$ is its complex amplitude²⁵. A similar approach was taken by Sanpera *et al.*²⁷, who divided the electron wavefunction into bound and continuum parts and modelled the continuum part as a gaussian wave packet. Provided that $a[k(\omega)]$ is known, then measuring the harmonic spectrum—its amplitude, phase and polarization—is equivalent to measuring this integral, evaluated for different $k(\omega)$. This is an experimental determination of the one-dimensional spatial Fourier transform of $\mathbf{r}\Psi_g(\mathbf{r})$.

Experimentally, we concentrate on a specific molecule, N_2 . High

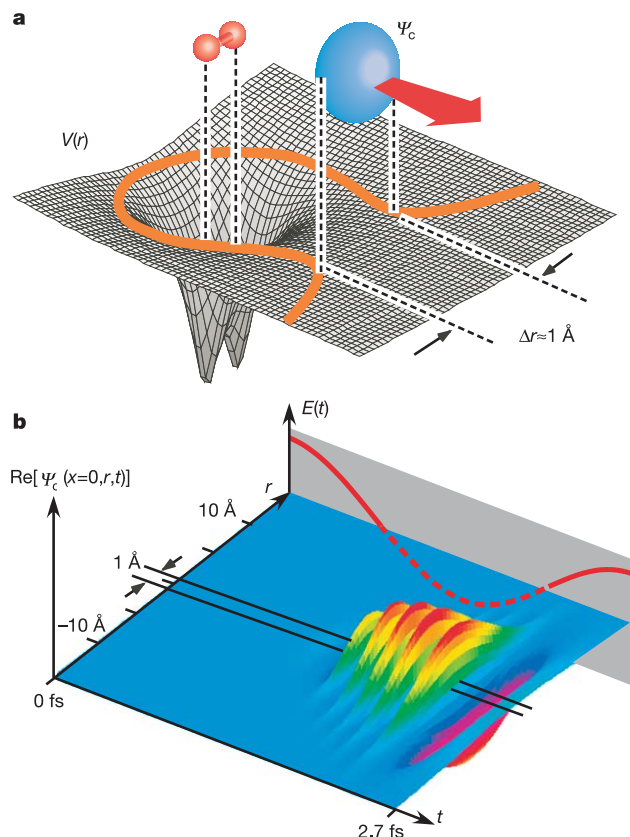


Figure 1 Illustration of the tunnel ionization process from an aligned molecule. **a**, The orange line on the potential surface is an isopotential contour slightly above the energy level of the bound state at the peak of the field amplitude. The opening of the contour shows the saddle point region where the bound-state electron wavefunction will tunnel through to the continuum. The lateral spread of the electron wave packet Ψ_c is determined by the width of the saddle point region Δr that depends on the molecular alignment. The wave packet expands in the lateral direction during propagation in the continuum because of the initial momentum spread Δp given by the uncertainty principle, $\Delta p \approx \hbar/\Delta r$. **b**, Illustration of the re-colliding wave packet seen by a molecule (the real part of the wave packet $\text{Re}[\Psi_c(x=0, t)]$ is shown). The kinetic energy at the time of re-collision is taken from the classical trajectory, and determines the instantaneous frequency of the wave packet as seen by the molecule. The lateral spread is calculated by the free expansion of a gaussian wave packet with an initial $1/e$ full width of 1 Å.

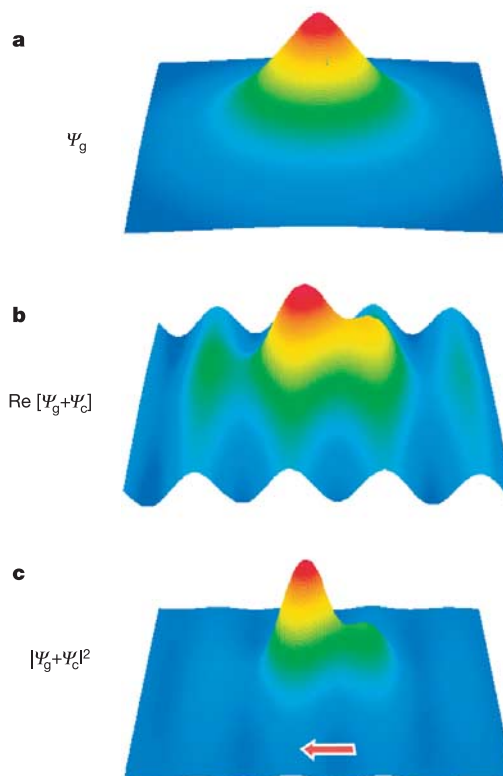


Figure 2 Illustration of a dipole induced by the superposition of a ground-state wavefunction Ψ_g and a re-colliding plane wave packet Ψ_c . **a**, Bound-state wavefunction (for example, H atom, 1 s). **b**, The real part of the superposition of the bound-state wavefunction Ψ_g and the continuum plane wave Ψ_c . **c**, The total electron density distribution $|\Psi_g + \Psi_c|^2$. The superposition of the two wavefunctions induces a dipole $d(t)$ as shown by the red arrow. As the wave packet propagates, the induced dipole oscillates back and forth and leads to the emission of harmonics.

harmonics are generated with a 30-fs, $2 \times 10^{14} \text{ W cm}^{-2}$, 800-nm, horizontally polarized probe pulse in a $\sim 10^{17} \text{ cm}^{-3}$ molecular gas jet described in the Methods section. The molecular alignment axis is rotated relative to the polarization of the probe pulse.

The probe pulse intensity was kept as low as possible to avoid saturation or depletion of the ground state. The molecular beam thickness was less than 1 mm and the laser focus was before the beam—conditions that minimize phase mismatch and select only the short electron trajectories²⁸.

Figure 3 shows the intensities of each harmonic as the laser polarization is rotated from 0 to 90° (in steps of 5°) with respect to the molecular axis. The intensity contrast between 0 and 90° exceeds an order of magnitude for some harmonics. We use these data to reconstruct the molecular orbital of N₂. Also shown is the spectrum from the reference atom, argon, recorded under conditions identical to those for the nitrogen.

Calibrating the continuum wave packet

As discussed above, the harmonic spectrum is an experimental evaluation of the dipole, $\mathbf{d}(\omega)$. If we could evaluate the plane-wave amplitude $a[k(\omega)]$ independently, then our measurement would determine $\int \psi_g(\mathbf{r})(\mathbf{e} \cdot \mathbf{r}) \exp[ik(\omega)x] d\mathbf{r}$ —that is, the spatial Fourier components of $\mathbf{r}\psi_g(\mathbf{r})$. One way to do this is to perform the same experiment with a reference atom.

Argon is very similar to N₂ in its response to strong laser fields, having nearly the same ionization potential and intensity-dependent ionization probability⁶. This is confirmed by the dependence of the instantaneous ionization rates²⁹ for atoms, and for different orientations of N₂ (ref. 30). That means that the first, critical, step in the three-step high harmonic generation process is the same. Because the laser field dominates wave packet motion in the direction of the laser field, the second step, which determines the chirp of the re-colliding wave packets seen by Ar or N₂, will be the same. Thus, $a[k(\omega)]$ will be the same.

The continuum wave packet will also be similar for Ar and N₂. The narrow saddle point through which the electron tunnels acts as a spatial filter that removes much of the structure of the orbital from the continuum wave packet. This can be seen in numerical simulations³¹. By measuring the ellipticity dependence of the high harmonic signal²⁴ produced by N₂ and argon, we confirmed that the lateral spread of the wave packets is similar. The ionization rate of N₂ is angle-dependent^{30,32}, but is readily measured from the ion yield, and varies only by 25% for N₂ (ref. 33). This variation is almost cancelled by the angular dependence of the wave-packet

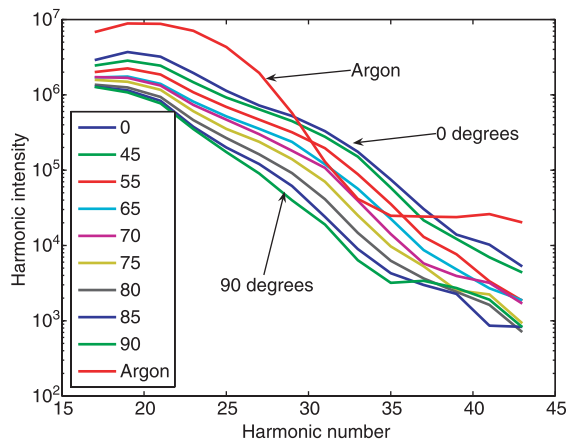


Figure 3 High harmonic spectra were recorded for N₂ molecules aligned at 19 different angles between 0 and 90° relative to the polarization axis of the laser. For clarity, only some of the angles have been plotted above. The high harmonic spectrum from argon is also shown; argon is used as the reference atom. Clearly the spectra depend on both the alignment angle and shape of the molecular orbital.

spread in the lateral direction²³. Thus we use the harmonic spectrum from argon to determine $|a[k(\omega)]|$ without including the angular dependence of tunnel ionization rate and wave-packet spread in the continuum.

Tomographic reconstruction of the orbital

We have shown that (1) tunnel ionization, owing to its nonlinearity, is extremely selective^{29,30}. In N₂, it ionizes only the highest electronic state—the HOMO. And (2), the harmonic spectra contain a range of spatial Fourier components of the shape of this single orbital. We have recorded the harmonic spectra at a series of angles between the molecular axis and the re-colliding electron. We now show that we can invert this information to obtain the shape of the orbital.

The Fourier slice theorem¹³ proves that the Fourier transform of a projection P_θ is equal to a cut at angle θ through the two-dimensional Fourier transform F of the object. This is the basis of computed tomography based on the inverse Radon transform. Our dipole is the Fourier transform of a projection of the wavefunction, and so can be inverted.

We describe the mathematical details of the tomographic reconstruction in the Methods section. This procedure can reconstruct orbital shapes with symmetries such as σ_g , π_g and π_u , using harmonics 17–51 of an 800-nm laser field and 25 angles from 0° to 180° (fewer angles are needed for symmetric molecules).

A complete inversion of a general orbital requires knowledge of the relative phase and amplitude of each harmonic for two ortho-

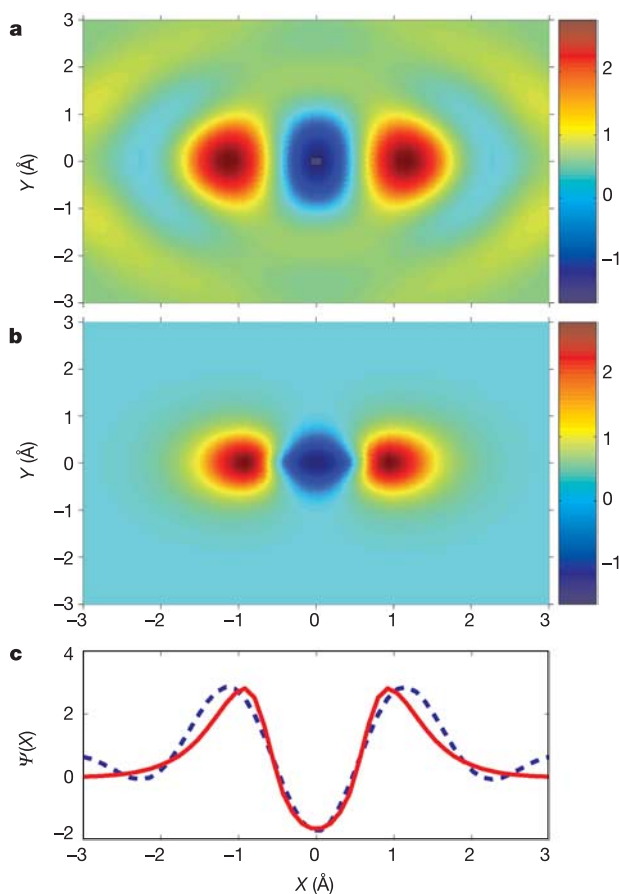


Figure 4 Molecular orbital wavefunction of N₂. **a**, Reconstructed wavefunction of the HOMO of N₂. The reconstruction is from a tomographic inversion of the high harmonic spectra taken at 19 projection angles. Both positive and negative values are present, so this is a wavefunction, not the square of the wavefunction, up to an arbitrary phase. **b**, The shape of the N₂ 2p σ_g orbital from an *ab initio* calculation. The colour scales are the same for both images. **c**, Cuts along the internuclear axis for the reconstructed (dashed) and *ab initio* (solid) wavefunctions.

gonal polarizations. Although we measured only the amplitude (the phase of each harmonic can be directly measured²⁸) the relative phase of the harmonics is known from first principles. As with any classical or quantum resonance, a phase jump of π occurs as the driving (electron) frequency moves across resonance. Calculations by Hay *et al.*⁹ show minimum harmonic signal and a π phase jump at the 25th harmonic for 0° when the re-collision electron wavelength resonates with the projected molecular dimension. We assume a phase jump of π where the measured dipole of N_2 goes through a minimum, corresponding to a change of sign of the dipole. This occurs at the 25th harmonic.

The reconstructed molecular orbital of N_2 is shown in Fig. 4a, based upon the 19 projections. It can be compared with the *ab initio* orbital calculation of the N_2 $2p$ σ_g orbital. Note that the recovered wavefunction has both positive and negative lobes, so it is not the square of the wavefunction, but the wavefunction itself, up to an arbitrary phase.

Our orbital reconstruction method is similar to medical tomography. However, we image a single orbital among many orbitals. The coherence of the re-collision provides this specificity. Of all the occupied or unoccupied states of the molecule, the re-colliding electron wave packet is only coherent with the state from which it tunnelled (or, as we will show below, with any other coherently related state). Measurable high harmonic emission requires a macroscopic number of ionizing molecules, so only the coherent emission interferes constructively and is observed experimentally. This coherent filtering is conceptually similar to the homodyne detection technique and is naturally implemented in harmonic generation.

Tomographic approaches have previously been used for molecular measurements, but in a very different way^{34,35}. Although tomographic reconstruction algorithms are used, they are based on 'rotations' in the position-momentum phase-space distributions, not physical rotations of an object. These techniques yield the one-dimensional Wigner distribution $W(x,p)$ of the phase-space density of a vibrating³⁴ or dissociating³⁵ molecule. In contrast, we measure the actual three-dimensional wavefunction of a single electron orbital of a molecule.

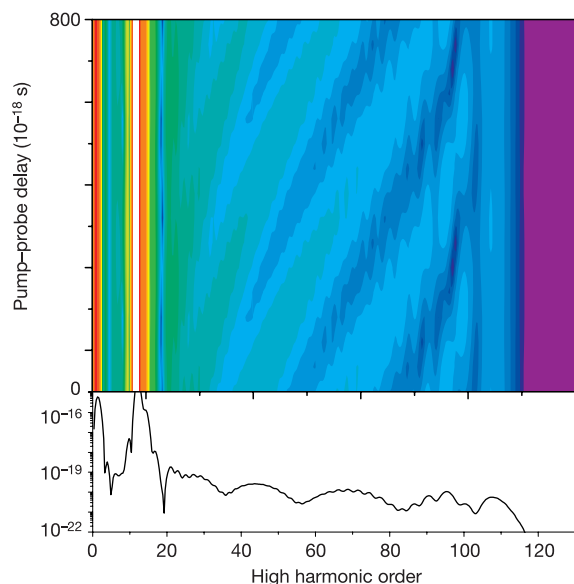


Figure 5 A one-dimensional Schrödinger calculation shows that attosecond electronic wave-packet motion is resolved in the high harmonic spectra. The bottom curve shows a spectrum at a particular pump-probe delay time; the minima are due to interference caused by the wave-packet motion. The top picture shows the spectra at a range of time delays, showing that the minima move with pump-probe time delay. The simulation populated the ground and first excited state (9.5 eV above ground) of a model atom to create an electronic wave packet.

Watching electrons move

So far, we have shown that we can take a snapshot of a molecular orbital using 30-fs laser pulses. Thus, in a pump-probe experiment, it is possible to image (probe) changes to molecular orbitals on the timescale of nuclear motion (the femtosecond timescale). However, electronic motion can be much faster. Can orbital imaging be extended to measuring electronic wave packets? We now show that the motion of this electronic wave packet can indeed be mapped to the harmonic spectrum.

It may seem that measuring wave-packet dynamics contradicts our assertion that the coherence of high harmonic generation selects a single orbital. However, an electronic wave packet is a coherent superposition of electronic states. Therefore, all states forming the wave packet are coherent with the returning electron provided it tunnelled from one or more of the states. They all contribute to the induced dipole that emits harmonics, even though only the highest-energy state contributed to the ionization. Therefore the relative intensity of each harmonic will depend on the relative phases between these electronic states.

To demonstrate how the bound-state wave-packet motion appears in harmonic spectra, we use the one-dimensional time-dependent Schrödinger equation to simulate an electron in a diatomic potential in the presence of an intense laser field. The field is chosen at a wavelength of 1.6 μm and duration of 8 fs to induce a single re-collision. To form a fast-moving bound-state wave packet, two electronic states with energy difference of $\Delta E = 9.5$ eV are equally populated before the arrival of the laser pulse. This energy difference corresponds to wave-packet motion with a period of 435 as. By varying the initial phase of the two states, we perform a numerical pump-probe measurement of the wave-packet motion.

Figure 5 shows the calculated harmonic spectrum versus the initial phase of the two states. The structure in the harmonic spectrum synchronizes with the bound-state wave-packet motion. Both the modulation in the single time delay spectrum and the movement of the modulation in the pump-probe spectrum measure wave-packet dynamics.

Promises and challenges

Our experimental condition— N_2 illuminated by an 800-nm pulse at an intensity of $2 \times 10^{14} \text{ W cm}^{-2}$ —is not special. Ionization intensities of $\sim 10^{14} \text{ W cm}^{-2}$ are typical for small- to intermediate-sized molecules³⁶. For these laser parameters a re-collision electron has a wavelength of 1–2 Å—the correct range for measuring orbital structure. Thus, it appears feasible to extend tomography to many other molecules, provided there is a general method for determining $a[k(\omega)]$. This seems very promising. The phase dependence of the ionization rate determines $a[k(\omega)]$ up to a global constant describing the total ionization probability. Theories of time-dependent ionization of atoms²⁹ and small molecules³⁰ shows that the phase-dependent ionization rate is determined mainly by the ionization potential. If this remains true for more-complex molecules, then a reference atom (or molecule) can always be found to experimentally determine $a[k(\omega)]$. If a reference atom is not possible then $a[k(\omega)]$ can be characterized experimentally through the ellipticity dependence of the harmonics²⁴, and by directly measuring the electron spectrum in elliptically or circularly polarized light.

But does the strong electric field of the laser modify the orbital? In small molecules the modification is small, because the re-colliding electron returns near the zero-crossing of the optical field. However, it may not be small in large molecules whose bound electrons can move more freely: that is, where the molecular polarizability is larger. Thus, it may be necessary to 'shield' the orbital from the intense field. One method of shielding is to exploit traditional spectroscopy. A state, first excited by resonant light, can provide the

continuum electron needed for tomography as long as it remains coherent with the ground state. Because the excited state will ionize much more easily than the ground state, tomography can be performed at field strengths in which the ground-state wavefunction is unperturbed. (Longer laser wavelengths will be required to keep the electron kinetic energy high in order to maintain sufficient spatial resolution.) Exploiting excited states has two additional advantages. First, the direction of the transition moment selects the alignment of those molecules that will ultimately ionize. Second, any orbital that is optically accessible can be imaged, not just the HOMO.

The ability of harmonics to measure the three-dimensional structure of orbitals is most important for molecular dynamics. It should be possible to observe an orbital while it changes its symmetry as a result of a curve-crossing, or to observe a chemical bond being broken—the very foundations of chemistry. With our wave-packet approach to align molecules, the only restriction is that the molecular dynamics must be faster than the duration of a rotational revival. To observe electronic rearrangement during chemical dynamics we would need to add an additional pulse to our pulse sequence³⁷. Its function would be to induce dynamics. The first pulse aligns the molecule; the second one initiates the vibrational (or electronic) dynamics that we wish to study; and the final pulse would produce the harmonic radiation that could probe the evolution of the wavefunction. □

Methods

Experimental details

For all measurements, we used a pulsed gas jet with a density of $\sim 10^{17} \text{ cm}^{-3}$ in the interaction region. The laser beams were focused $\sim 1 \text{ mm}$ below the orifice of the gas nozzle. The intensity of the pump pulse was set at $\sim 4 \times 10^{13} \text{ W cm}^{-2}$. This intensity is below the ionization threshold, where no ions are detected by a d.c.-biased electrode placed near the gas jet, and no harmonic emission is detected. The intensity of the probe pulse was estimated to be $\sim 2 \times 10^{14} \text{ W cm}^{-2}$ from the cut-off of high harmonics. This laser intensity is well below the saturation of ionization and harmonic generation, ensuring that high harmonics are emitted mostly at the peak of the laser pulse without depleting the ground state.

Tomographic procedure

The angular frequency ω of the radiated extreme-ultraviolet field, and the wavenumber k corresponding to the electron wave that produced it at the time of recollision, are related by $k(\omega) = (2\omega)^{1/2}$ in atomic units. The transition dipole moment between the orbital wavefunction and the continuum electron is $\mathbf{d}(\omega; \theta) = \langle \psi(\mathbf{r}; \theta) | \mathbf{r} | \exp[ik(\omega)x] \rangle$. This is a complex vector, and $\psi(\mathbf{r}; \theta)$ represents the orbital wavefunction rotated by Euler angle θ . Assuming perfect phase-matching, the extreme-ultraviolet signal that is emitted is given as $S(\omega; \theta) = N^2(\theta) \omega^4 |a[k(\omega)] \mathbf{d}(\omega; \theta)|^2$. Here $a[k(\omega)]$ is the complex amplitude of component k of the continuum wave packet, and $N(\theta)$ is the number of ions produced.

The value of $a[k(\omega)]$ is determined experimentally by recording $S(\omega)$ for a reference argon atom whose orbital ($2p_x$), and hence $d(\omega)$, is known. Hence, $a[k(\omega)] = \omega^{-2} [S_{\text{ref}}(\omega)]^{1/2} / [\langle \psi_{\text{ref}} | r | k \rangle]^2$. This calibration not only characterizes the continuum wave packet, it also includes experimental factors such as detector efficiency.

Then $S(\omega; \theta)$ is recorded for an unknown molecule, yielding its transition dipole moment $d(\omega; \theta)$. The definition of $d(\omega; \theta)$ is a spatial Fourier transform of the orbital in the x direction. We apply the Fourier slice theorem¹³ to do the inversion, with one important modification. The Fourier transform contains the factor $\mathbf{r}$, which, being defined in the laboratory frame, does not rotate with the molecular frame, θ .

The tomographic inversion requires two intermediate functions,

$$f_x(x, y) = \sum_{\theta} \sum_{\omega} [d_x(\omega; \theta) \cos \theta + d_y(\omega; \theta) \sin \theta] \exp[ik(\omega)(x \cos \theta + y \sin \theta)]$$

$$f_y(x, y) = \sum_{\theta} \sum_{\omega} [-d_x(\omega; \theta) \sin \theta + d_y(\omega; \theta) \cos \theta] \exp[ik(\omega)(x \cos \theta + y \sin \theta)]$$

Then the wavefunction is given as:

$$\psi(x, y) = \text{Re}(f_x/x + f_y/y).$$

We must also include the angle dependence of the ionization rate to normalize each spectrum. The ionization rate is known for simple molecules such as N_2 (refs 30, 32) but can be measured at the same time as the harmonic spectrum is being recorded³³ simply by measuring the total ion yield, $N(\theta)$.

Received 8 June; accepted 9 November 2004; doi:10.1038/nature03183.

- Mulliken, R. S. Electronic structures of polyatomic molecules and valence. II. General considerations. *Phys. Rev.* **41**, 49–71 (1932).
- Linus, C. P. *The Nature of the Chemical Bond and the Structure of Molecules and Crystals* (Cornell Univ. Press, Ithaca, New York, 1960).
- Brion, C. E., Cooper, G., Zheng, Y., Litvinyuk, I. V. & McCarthy, I. E. Imaging of orbital electron

- densities by electron momentum spectroscopy—a chemical interpretation of the binary (e, 2e) reaction. *Chem. Phys.* **270**, 13–30 (2001).
- Binning, G., Rohrer, H., Gerber, Ch. & Weibel, E. Surface studies by scanning tunneling microscopy. *Phys. Rev. Lett.* **49**, 57–61 (1982).
- Sakai, H. & Miyazaki, K. High-order harmonic generation in nitrogen molecules with subpicosecond visible dye-laser pulses. *Appl. Phys. B* **61**, 493–498 (1995).
- Liang, Y., Augst, A., Chin, S. L., Beaudoin, Y. & Chaker, M. High harmonic generation in atomic and diatomic molecular gases using intense picosecond laser pulses—a comparison. *J. Phys. B* **27**, 5119–5130 (1994).
- Velotta, R., Hay, N., Manson, M. B., Castillejo, M. & Marangos, J. P. High-order harmonic generation in aligned molecules. *Phys. Rev. Lett.* **87**, 183901 (2001).
- de Nalda, R. *et al.* Role of orbital symmetry in high-order harmonic generation from aligned molecules. *Phys. Rev. A* **69**, 031804(R) (2004).
- Hay, N. *et al.* Investigations of electron wave-packet dynamics and high-order harmonic generation in laser-aligned molecules. *J. Mod. Opt.* **50**, 561–571 (2003).
- Hay, N. *et al.* High-order harmonic generation in laser-aligned molecules. *Phys. Rev. A* **65**, 053805 (2002).
- Lein, M., Corso, P. P., Marangos, J. P. & Knight, P. L. Orientation dependence of high-order harmonic generation in molecules. *Phys. Rev. A* **67**, 023819 (2003).
- Parr, R. G. & Yang, W. *Density-functional Theory of Atoms and Molecules* (Oxford Univ. Press, New York, 1989).
- Kak, A. C. & Slaney, M. *Principles of Computerized Tomographic Imaging* (Society for Industrial and Applied Mathematics, New York, 2001).
- Stapelfeldt, H. & Seideman, T. Aligning molecules with strong laser pulses. *Rev. Mod. Phys.* **75**, 543–557 (2003).
- Stapelfeldt, H., Sakai, H., Constant, E. & Corkum, P. B. Deflection of neutral molecules using the nonresonant dipole force. *Phys. Rev. Lett.* **79**, 2787–2790 (1997).
- Sakai, H. *et al.* Controlling the alignment of neutral molecules by a strong laser field. *J. Chem. Phys.* **110**, 10235–10238 (1999).
- Larsen, J. J., Hald, K., Bjerre, N. & Stapelfeldt, H. Three dimensional alignment of molecules using elliptically polarized laser fields. *Phys. Rev. Lett.* **85**, 2470–2473 (2000).
- Sakai, H., Minemoto, S., Nanjo, H., Tanji, H. & Suzuki, T. Controlling the orientation of polar molecules with combined electrostatic and pulsed nonresonant laser fields. *Phys. Rev. Lett.* **90**, 083001 (2003).
- Rosca-Pruna, F. & Vrakking, M. J. J. Experimental observation of revival structures in picosecond laser-induced alignment of I_2 . *Phys. Rev. Lett.* **87**, 153902 (2001).
- Dooley, P. W. *et al.* Direct imaging of rotational wave-packet dynamics of diatomic molecules. *Phys. Rev. A* **68**, 023406 (2003).
- Delone, N. B. & Krainov, V. P. *Multiphoton Processes in Atoms* (Springer, Heidelberg, 2000).
- Corkum, P. B. Plasma perspective on strong-field multiphoton ionization. *Phys. Rev. Lett.* **71**, 1994–1997 (1993).
- Niikura, H. *et al.* Sub-laser-cycle electron pulses for probing molecular dynamics. *Nature* **417**, 917–922 (2002).
- Dietrich, P., Burnett, N. H., Ivanov, M. & Corkum, P. B. High-harmonic generation and correlated two-electron multiphoton ionization with elliptically polarized light. *Phys. Rev. A* **50**, R3585–R3588 (1994).
- Lewenstein, M., Balcou, Ph., Ivanov, M. Yu., L’Huillier, A. & Corkum, P. B. Theory of high-harmonic generation by low-frequency laser fields. *Phys. Rev. A* **49**, 2117–2132 (1994).
- Burnett, K., Reed, V. C., Cooper, J. & Knight, P. L. Calculation of the background emitted during high-harmonic generation. *Phys. Rev. A* **45**, 3347–3349 (1992).
- Sanpera, A. *et al.* Can harmonic generation cause non-sequential ionization? *J. Phys. B* **31**, L841–L848 (1998).
- Mairesse, Y. *et al.* Attosecond synchronization of high-harmonic soft x-rays. *Science* **302**, 1540–1543 (2003).
- Yudin, G. & Ivanov, M. Yu. Nonadiabatic tunnel ionization: Looking inside a laser cycle. *Phys. Rev. A* **64**, 013409 (2001).
- Tong, X. M., Zhao, Z. X. & Lin, C. D. Theory of molecular tunneling ionization. *Phys. Rev. A* **66**, 033402 (2002).
- Spanner, M., Smirnova, O., Corkum, P. B. & Ivanov, M. Yu. Reading diffraction images in strong field ionization of diatomic molecules. *J. Phys. B* **37**, L243–L250 (2004).
- Otobe, T., Yabana, K. & Iwata, J.-I. First-principles calculations for the tunnel ionization rate of atoms and molecules. *Phys. Rev. A* **69**, 053404 (2004).
- Litvinyuk, I. V. *et al.* Alignment-dependent strong field ionization of molecules. *Phys. Rev. Lett.* **90**, 233003 (2003).
- Dunn, T. J., Wamsley, I. A. & Mukamel, S. Experimental determination of the quantum mechanical state of a molecular vibrational mode using fluorescence tomography. *Phys. Rev. Lett.* **74**, 884–887 (1995).
- Skovsen, E., Stapelfeldt, H., Juhl, S. & Mølmer, K. Quantum state tomography of dissociating molecules. *Phys. Rev. Lett.* **91**, 090406 (2003).
- Hankin, S. M., Villeneuve, D. M., Corkum, P. B., & Rayner, D. M. Nonlinear ionization of organic molecules in high intensity laser fields. *Phys. Rev. Lett.* **84**, 5082–5085 (2000).
- Lee, K. F. *et al.* Two-pulse alignment of molecules. *J. Phys. B* **37**, L43–L48 (2004).

Acknowledgements In addition to the NRC, we acknowledge financial support from the National Science and Engineering Research Council, Photonic Research Ontario, the Canadian Institute for Photonic Innovation, the Alexander von Humboldt-Stiftung and the Japan Society for the Promotion of Science. We thank M. Yu. Ivanov, M. Spanner, J. P. Marangos, M. Lein, P. H. Bucksbaum, I. A. Wamsley, D. Jonas, J. Tse and J. G. Underwood for discussions.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.M.V. (david.villeneuve@nrc.ca).

Structural basis for the assembly of a nuclear export complex

Yoshiyuki Matsuura & Murray Stewart

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

The nuclear import and export of macromolecular cargoes through nuclear pore complexes is mediated primarily by carriers such as importin- β . Importins carry cargoes into the nucleus, whereas exportins carry cargoes to the cytoplasm. Transport is orchestrated by nuclear RanGTP, which dissociates cargoes from importins, but conversely is required for cargo binding to exportins. Here we present the 2.0 Å crystal structure of the nuclear export complex formed by exportin Cse1p complexed with its cargo (Kap60p) and RanGTP, thereby providing a structural framework for understanding nuclear protein export and the different functions of RanGTP in export and import. In the complex, Cse1p coils around both RanGTP and Kap60p, stabilizing the RanGTP-state and clamping the Kap60p importin- β -binding domain, ensuring that only cargo-free Kap60p is exported. Mutagenesis indicated that conformational changes in exportins couple cargo binding to high affinity for RanGTP, generating a spring-loaded molecule to facilitate disassembly of the export complex following GTP hydrolysis in the cytoplasm.

Macromolecules are exchanged between the cytoplasmic and nuclear compartments through nuclear pore complexes (NPCs), huge macromolecular assemblies that perforate the nuclear envelope¹. Specific carrier molecules, many of which belong to the karyopherin- β superfamily, mediate transport of cargoes through NPCs and also have an emerging role as regulators of other cellular processes, including mitosis and gene expression². Frequently, transport is controlled by the nucleotide state of the Ras-family GTPase Ran that modulates the affinity of carriers for their cargoes through complex formation. In nuclear protein import, for example, cargoes containing a nuclear localization sequence (NLS) most commonly bind the carrier importin- β (Kap95p in yeast) via the importin- α (Kap60p) adaptor in the cytoplasm, where Ran is predominantly in the GDP-bound state. After passage through NPCs, cargo is released by nuclear RanGTP that complexes with Kap95p, displacing Kap60p. The importin- β -binding (IBB) domain of Kap60p then displaces the NLS, releasing the cargo³. Kap95p and Kap60p are then recycled to the cytoplasm to participate in further transport cycles. Kap95p returns complexed with RanGTP, whereas Kap60p export is mediated by the karyopherin- β family member, CAS (Cse1p), also complexed with RanGTP^{4–6}. Finally, RanGTP hydrolysis, stimulated by cytoplasmic RanGAP and RanBP1, disassembles the Kap95p:RanGTP and Cse1p:Kap60p:RanGTP complexes^{1–6}. Analogous cycles are thought to operate for the transport of many other cargoes by other karyopherin- β -family members, but paradoxically, whereas RanGTP functions to release import cargoes in the nucleus, it is required for export carriers to bind their cargoes^{1,2}.

The structure of importin- β ⁷, Kap- β ²⁸ and several of their complexes^{9–11} has given an appreciation of how RanGTP functions to release cargo in nuclear protein import, but the basis for the opposite function of RanGTP in nuclear export, especially why complex formation is required to bind rather than release cargo in the nucleus, has remained obscure. Here we have addressed this question by establishing the crystal structure of a nuclear export complex. In addition to defining the precise interface between an export carrier and its cargo, the structure of the Cse1p:Kap60p:RanGTP complex shows how RanGTP binding is intimately associated with cargo binding and indicates that it enables energy to be stored that facilitates cargo release when the RanGTP is hydrolysed in the cytoplasm.

Cse1p:Kap60p:RanGTP complex

We obtained crystals of the Cse1p:Kap60p:RanGTP complex using full-length Cse1p, Kap60p residues 1–530 and RanGTP residues 1–176 (removal of residues 177–216 favours Ran adopting the GTP-bound conformation¹²). Molecular replacement using Ran and Kap60p as models enabled building of the Cse1p chain and, after refinement, gave a model for the complex with an R-factor of 23.4% (R_{free} 26.8%) and good stereochemistry (Table 1). Although the structure of Cse1p superficially resembled importin- β ⁷ and Kap- β ²⁸, there were several crucial differences that were important functionally. Cse1p is constructed from a right-handed superhelix of 20 HEAT repeats (with two long inserts in HEAT8 and HEAT19; see Supplementary Fig. 1) but does not have the large loops and extended HEAT helices that are important for cargo recognition or RanGTP-mediated cargo release in importin- β ^{7,10} and Kap- β ²⁸. The long insert between the outer and inner helices of HEAT19 in Cse1p was remarkable in extending back along the concave inner surface of the molecule almost to its mid-point and binding both RanGTP and Kap60p. The central region of the Cse1p superhelix (HEAT8–12)

Table 1 Crystallographic statistics

Data collection statistics	
Space group	C2
Unit cell dimensions	$a = 272.68 \text{ Å}$, $b = 72.26 \text{ Å}$, $c = 97.64 \text{ Å}$, $\beta = 92.57^\circ$
Resolution range (Å)	20–2.0 (2.11–2.00)*
Wilson B-factor	31.7
Total observations	318,129 (30,129)*
Unique reflections	120,581 (16,799)*
Completeness (%)	94.3 (90.4)*
R_{merge} (%)	8.9 (39.6)*
I/σ	5.5 (1.8)*
Refinement statistics	
Number of reflections (working, test)	114,491, 6,089
$R_{\text{cryst}}/R_{\text{free}}$ (%)	23.4/26.8
Total number of non-H atoms	1,3081
Number of water molecules	486
r.m.s. deviation from ideal bond length (Å)	0.008
r.m.s. deviation from ideal bond angles (degrees)	1.361
Ramachandran plot (%)	
Core region	91.2
Allowed region	7.4
Generously allowed region	1.3
Disallowed region	0.1

*Parentheses refer to final resolution shell

was slightly twisted, generating two roughly orthogonal arches. The central twist appeared to be sustained, in part, by the HEAT8 insert packing against the inner surface of HEAT9–12. The conformation of the major portion of Kap60p that is constructed from Armadillo (ARM) repeats was essentially unaltered in the complex (root-mean-square deviation, r.m.s.d., 1.4 Å) and the IBB domain was bound along its inner surface in contact with the two NLS-binding sites, analogous to that seen with importin- α^3 . The structure of Ran in the Cse1p:Kap60p:RanGTP complex was essentially identical (r.m.s.d. 0.4 Å) to that in the importin- β :RanGMPNP complex⁹, and the density map clearly showed that it was bound to GTP (see below), indicating the nucleotide was not hydrolysed during crystallization.

The way in which Cse1p, Kap60p and RanGTP were arranged in the export complex was quite different to that seen in nuclear import complexes. In the export complex, both RanGTP and the carboxy-terminal region of Kap60p (ARM8–10) were clamped between the amino- and C-terminal arches of Cse1p (Fig. 1), with direct contacts also being made between RanGTP and the last Kap60p ARM repeat. The extensive interaction between Cse1p and the C-terminal region of Kap60p was consistent with the observation that C-terminal deletion of importin- α abolished binding to CAS¹³. The N-terminal part of the Kap60p ARM domain (ARM1–7) protruded outward from Cse1p, whereas the IBB domain folded back towards Cse1p, binding along the major and minor NLS binding sites on the Kap60p ARM repeats, as well as interacting directly with Cse1p on the outer surface of its N-terminal arch. The insert in Cse1p HEAT19 was involved intimately in these interactions. The interaction interfaces between the three molecules were remarkably extensive, with a total buried surface area of 9,410 Å² (Table 2), consistent with the nM dissociation constant of the complex⁶. The binding appeared to be highly co-operative, so that, rather than recognizing only a short stretch of residues as an export signal, Cse1p and RanGTP recognize both the

Table 2 Buried interfacial area

	Kap60p IBB	Kap60p ARM repeats	RanGTP
Cse1p	2,151 Å ²	2,982 Å ²	3,570 Å ²
Kap60p IBB		3,191 Å ²	0 Å ²
Kap60p ARM repeats			707 Å ²

extended polypeptide (the IBB domain) and also the large folded ARM domain of Kap60p. It is likely that formation of the ternary complex involved a series of co-ordinated steps so that the complex was effectively 'zipped up'.

Function of the Kap60p IBB domain

In addition to facilitating binding to Kap95p (importin- β), the IBB domain also functions in NLS release by competing for the same sites on Kap60p (importin- α)^{3,14} once it is released from Kap95p by RanGTP. In the Cse1p:Kap60p:RanGTP complex, the IBB autoinhibitory region (KRRNE, residues 54–58) binds the Kap60p major NLS-binding site (Fig. 2a, e) in a similar way to that seen in mouse importin- α^3 , while another IBB domain basic cluster (RRRR, residues 33–36) binds the minor NLS-binding site (Fig. 2b, e). Notably, residues 40–45 (QVELRK) between the two basic IBB clusters bind to the outer surface of Cse1p HEAT5–7 (Fig. 2c, e), and the IBB N terminus (FVPEYRR, residues 12–18) binds Cse1p HEAT2–4 (Fig. 2d, e). Cse1p and RanGTP hold the Kap60p C terminus so that the inner surface of the Kap60p ARM repeats faces towards the outer surface of the N-terminal arch of Cse1p on which the IBB binding sites are located. By clamping the IBB to the NLS binding sites, its autoinhibitory function would be reinforced. Moreover, the requirement for the IBB domain to be locked onto the ARM repeats for Cse1p to bind would ensure that only cargo-free Kap60p is exported.

Study of IBB mutants underscored the importance of interactions involving the IBB domain. R44D-Kap60p markedly weakened the

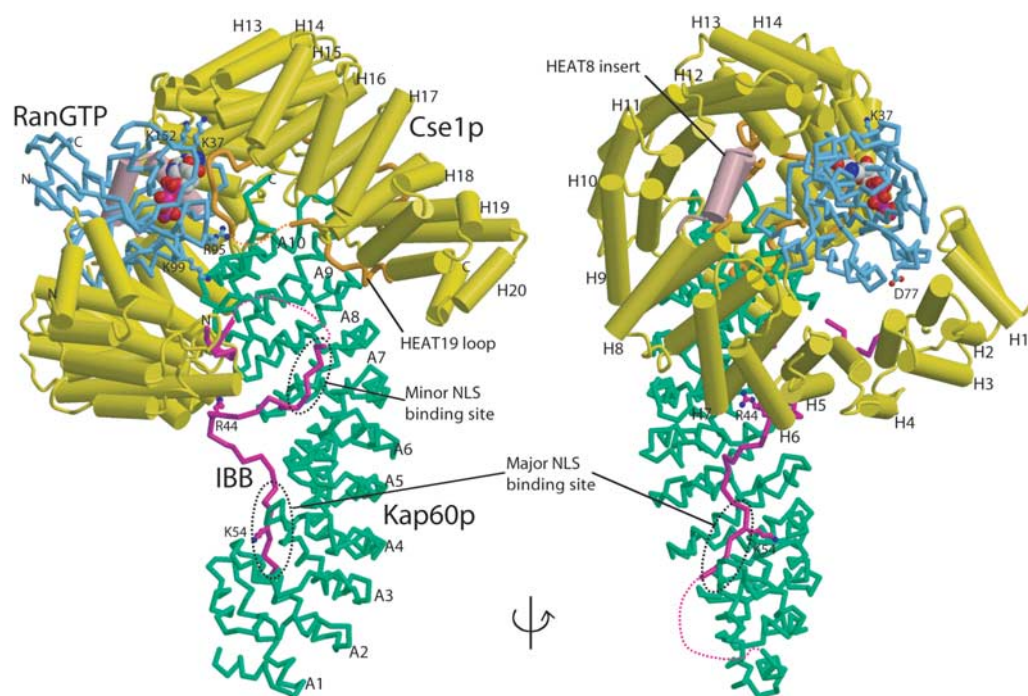


Figure 1 Two orthogonal views, showing an overview of the structure of the Cse1p:Kap60p:RanGTP complex. The figure illustrates how Cse1p (yellow) envelopes RanGTP (blue), and the C-terminal region of Kap60p (green) and its IBB domain (magenta). The Cse1p α -helices are represented by cylinders, whereas the HEAT8 insert

and HEAT19 loop are highlighted in pink and orange, respectively. The ARM repeats of Kap60p and HEAT repeats of Cse1p are labelled A1–A10 and H1–H20, respectively. GTP is shown as spacefilling spheres. Key residues mentioned in the text are shown in ball-and-stick representation.

Cse1p:Kap60p:RanGTP binding (Fig. 2f), confirming that the IBB:Kap60p interaction is crucial for assembly of the ternary complex. K54A-Kap60p (which has reduced autoinhibitory activity¹⁴) also weakened the binding (Fig. 2f), suggesting that it is only when the IBB binds to the NLS binding site that Kap60p binds Cse1p and RanGTP tightly. This would be an effective mechanism to discriminate between NLS-cargo bound and unbound Kap60p,

ensuring that only cargo-free Kap60p is committed to nuclear export. Consistent with this idea, the affinity of wild-type Kap60p for a NLS was weakened by Cse1p and RanGTP, whereas the NLS binding of both R44D- and K54A-Kap60p was insensitive to Cse1p and RanGTP (Fig. 2g). The same mechanism of NLS competition could be used by CAS and RanGTP to release mitotic spindle assembly factors such as TPX2 from importin- α during cell

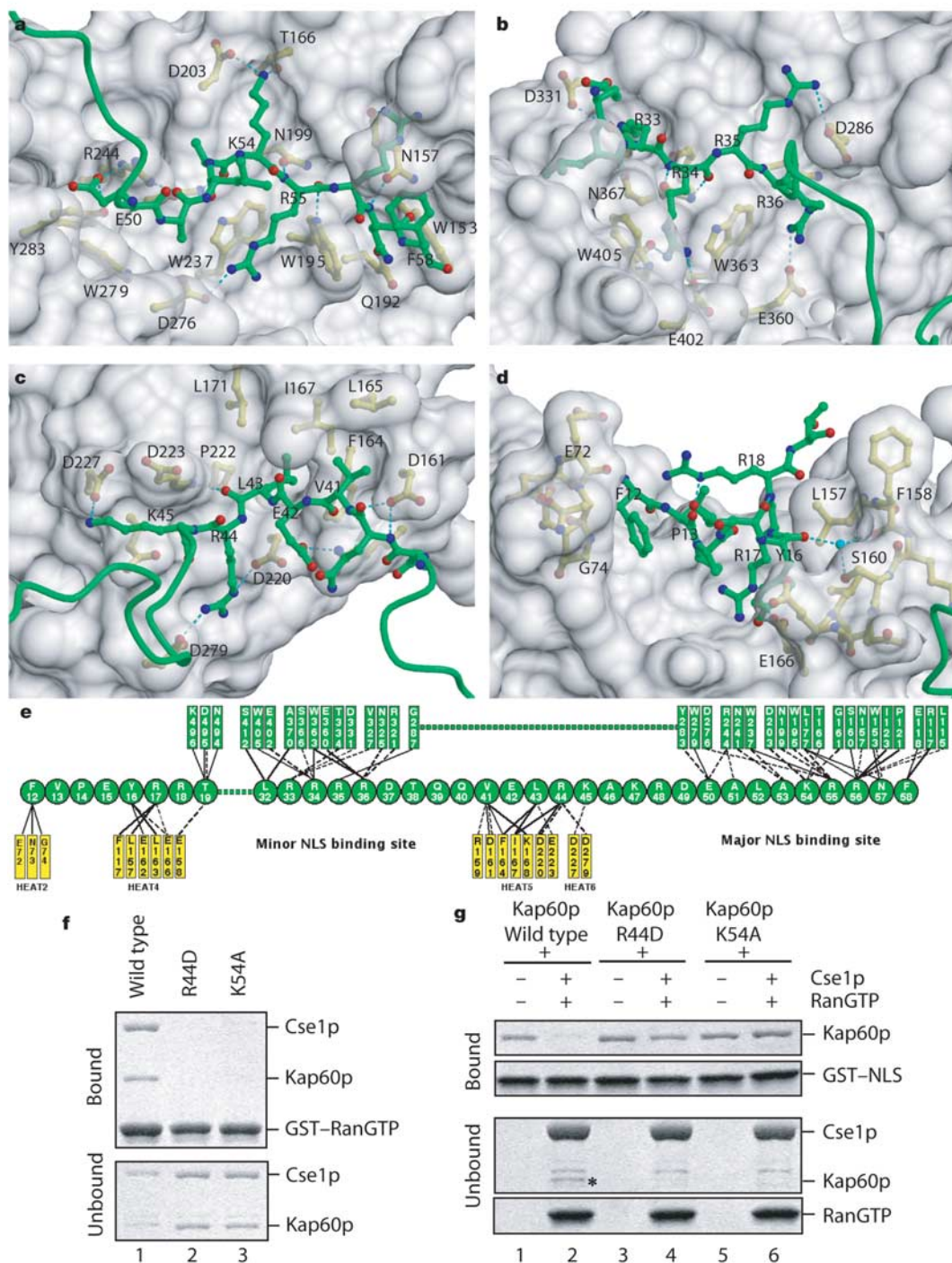


Figure 2 Role of the IBB domain in nuclear export complex formation. **a–d**, Interactions involving the Kap60p IBB domain (green) with key residues (shown under the transparent surface) on Kap60p (**a, b**) or Cse1p (**c, d**). **a**, Major NLS binding site. **b**, Minor NLS binding site. **c, d**, IBB binding sites on Cse1p HEAT5–7 (**c**) and Cse1p HEAT2–4 (**d**). **e**, Schematic illustration of interactions involving residues in the IBB domain (circles) with Kap60p (green rectangles) and Cse1p (yellow rectangles). Full lines represent hydrophobic interactions, broken lines H-bonds and salt-bridges. **f**, IBB mutations impair the assembly

of the ternary complex. Immobilized GST–RanGTP was treated with Cse1p and Kap60p. Whereas wild-type Kap60p bound, the R44D and K54A IBB mutants did not. **g**, IBB interactions are crucial for the competition with NLS. Immobilized GST–Cbp80p NLS was incubated with Kap60p, and after the unbound Kap60p was washed away, treated with buffer alone (lanes 1, 3, 5) or Cse1p:RanGTP (lanes 2, 4, 6). Although wild-type Kap60p (* in unbound fraction, lane 2) was released, the R44D and K54A IBB mutants remained bound to the NLS.

division¹⁵, and could account for some of the tumour suppressor and apoptotic functions of CAS¹⁶. The importance of the IBB:Cse1p interaction is also corroborated by the *cse1-1* (D220N) mutant¹⁷ that weakens the binding to Kap60p:RanGTP *in vitro*⁴, because Cse1p Asp220 forms a salt-bridge with Kap60p Arg44 (Fig. 2c, e). Moreover, in *cse1-1* cells, Kap60p accumulates in the nucleus, suggesting that the interaction is important for Kap60p export *in vivo*^{4,5}. The structure of the Kap60p ARM repeats in the Cse1p:Kap60p:RanGTP complex is nearly identical to the structure of Kap60p ARM repeat domain alone¹⁸ (r.m.s.d. 1.4 Å). Therefore, the binding of Cse1p and RanGTP does not induce allosteric conformational changes in the Kap60p ARM region, and it is the IBB domain that is primarily responsible for releasing NLS-cargoes.

RanGTP binds Cse1p at two sites

Cse1p interacts with RanGTP at two distinct sites (Fig. 3a). The first Cse1p Ran-binding site (N-interface) is formed by the inner surface of HEAT1–3 and packs against the Ran switch II loop. This interface is analogous to that formed between RanGMPNP and the importin-β N terminus⁹, consistent with the sequence similarity between importins and exportins being the highest in this region¹⁹. The second Ran-binding interface (C-interface) interacts with the region centred on the Ran switch I loop. It is formed by HEAT13–14 together with the long internal loop in HEAT19 and has no counterpart in importin-β. In addition to binding to switch I, the HEAT19 loop also packs over the guanine and ribose moieties of GTP, inhibiting nucleotide dissociation.

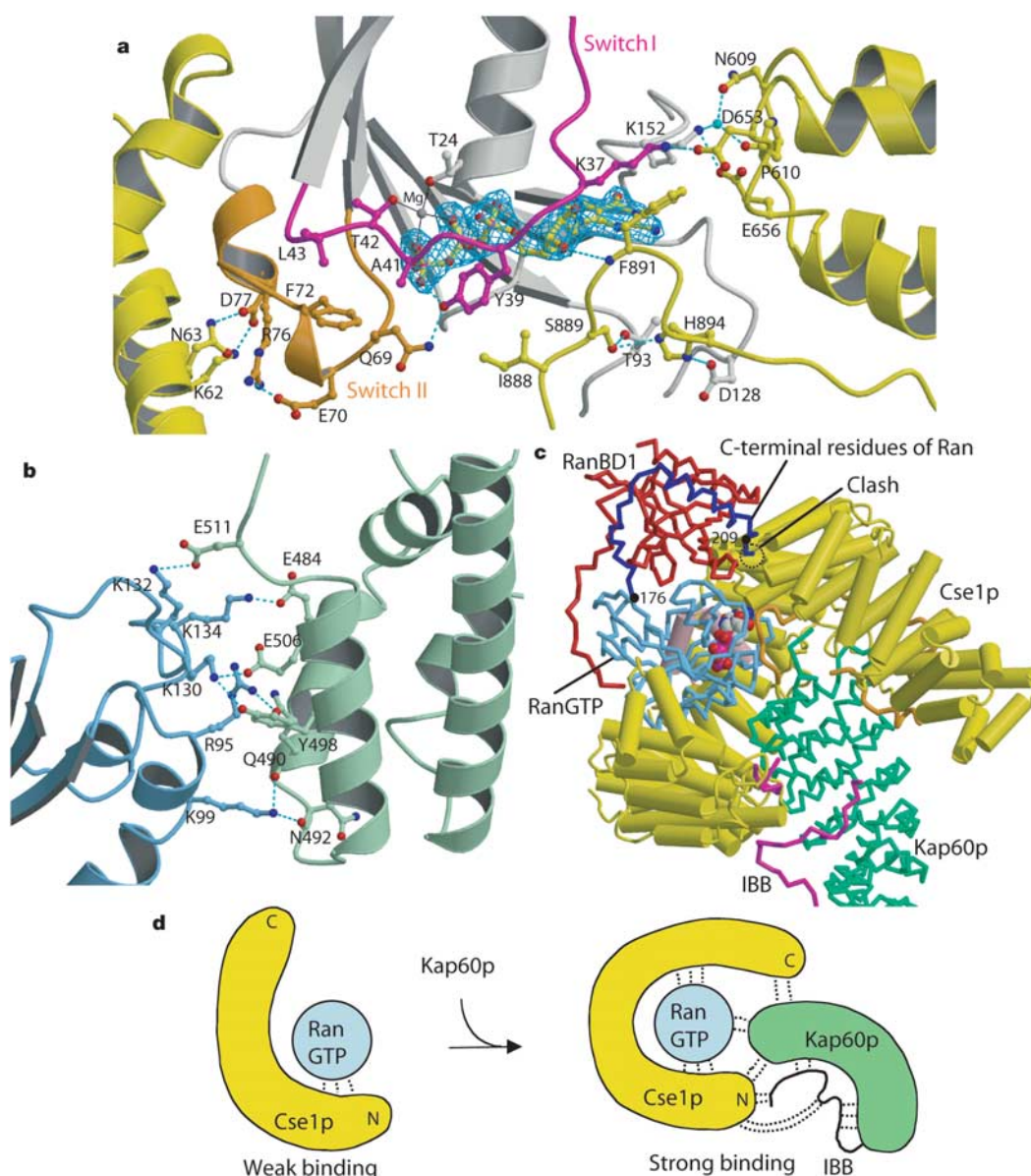


Figure 3 Role of Ran in the export complex. **a**, Interfaces between Cse1p (yellow) and Ran (switch I, magenta; switch II, orange). Ran Arg 76 and Asp 77 in the switch II loop are key components of the Cse1p N-interface, whereas Lys 37 (switch I loop) and Lys 152 bind the C-interface. An omit-annealed $F_o - F_c$ map around GTP at 3.5 Å is superimposed, confirming the nucleotide state. **b**, Ran Arg 95, Lys 99, Lys 130 and Lys 134 are key components of the interface between RanGTP (blue) and Kap60p (green). **c**, Docking RanBD1 onto the Cse1p:Kap60p:RanGTP complex shows a steric clash between Cse1p and the RanGTP C terminus. The coordinates of Ran bound to RanBD1 (red) were superimposed with Ran (residues 1–176) in the Cse1p:Kap60p:RanGTP complex.

C-terminal residues of Ran (176–210) in the Ran-RanBD1 complex, dark blue; other colour coding as Fig. 1. **d**, Model for Cse1p:Kap60p:RanGTP complex assembly. Dotted lines represent interactions. In the absence of Kap60p, Cse1p has a lower-energy conformation that binds Ran at only one site. When Kap60p binds, Cse1p is distorted into a higher-energy state that allows Ran to bind to the second site, generating a spring-loaded complex that would disassemble spontaneously on RanGTP hydrolysis. The representation of Cse1p here is highly schematic, and is intended only to imply a conformational change and not the precise structure of either state.

Clamping Ran at two sites, each of which changes conformation depending on the nucleotide state, would effectively lock it in the GTP-bound conformation. As observed in the importin- β :RanGMPNP complex⁹, the catalytic sidechain of Ran Gln 69 is flipped away from the GTP γ -phosphate, which would further inhibit GTP hydrolysis during nuclear export. Although importin- β and Kap- β ²⁸ also bind the basic patch on RanGTP (centred on Arg 140), this patch in the Cse1p:Kap60p:RanGTP complex interacts electrostatically with the acidic surface of ARM10 of Kap60p (Fig. 3b).

We used Ran mutants to explore the importance of these different interfaces. Ran mutations at the Cse1p N-interface (R76A/D77A) or C-interface (K37D/K152A), or at the Kap60p interface (R95D/K99A/K130A/K134A), markedly impaired formation of the Cse1p:Kap60p:RanGTP ternary complex (Fig. 4a). All of these Ran mutants bound normally to the yeast RanBP1 homologue Yrb1p (Fig. 4c), whose binding sites on Ran are different from Cse1p/Kap60p binding sites and instead are similar to the Ran-binding domains (RBD) of RanBP2²⁰, indicating that the overall RanGTP conformation was retained. The structure of the Cse1p:Kap60p:RanGTP complex also suggested how Yrb1p can itself disassemble that complex⁵. Superimposing the RanBD1:RanGMPNP²⁰ and Cse1p:Kap60p:RanGTP complex structures suggested that the Ran C-terminal DEDDDL motif (residues 211–216) would clash with the Cse1p C-interface if RanBP1 docked onto the complex (Fig. 3c). Yrb1p did not effectively compete with Cse1p and Kap60p when the C-terminal residues 210–216 of Ran were deleted (Fig. 4e), supporting the idea that the Ran C terminus is critical for export

complex disassembly and a further indication of the critical role of the Cse1p C-interface in export complex integrity.

Ran function in nuclear export

The extensive buried surface in the RanGTP:Cse1p interface (3,570 Å²) observed in the Cse1p:Kap60p:RanGTP complex is comparable to that in the RanGTP:importin- β (3,283 Å²) or RanGTP:Kap- β ²⁸ (4,027 Å²) interfaces, and the superhelix of these three karyopherin- β -family members follows a similar trajectory in the region bound to RanGTP (Supplementary Fig. 2). However, in contrast to importin- β , Cse1p does not bind RanGTP tightly in the absence of Kap60p⁵. Although the binding of RanGTP to Cse1p alone is weak, it is still detectable in pull-down assays using high protein concentrations. Ran mutations in the region that interacts with the Cse1p N-interface (R76A/D77A) weakened the Cse1p:RanGTP interaction so that it was no longer detectable, whereas mutations at the C-interface binding region (K37D/K152A) bound as strongly as wild type (Fig. 4d). This implied that, in the absence of Kap60p, Cse1p adopts a conformation different from that observed in the Cse1p:Kap60p:RanGTP complex and only the N-interface is available for RanGTP binding. If so, the energy of Kap60p binding could be used to distort Cse1p to a higher-energy, strained conformation so that the C-interface can also contact RanGTP (Fig. 3d). Most of the free energy of Kap60p:Cse1p, Kap60p:Ran and IBB:ARM interactions could then be used to overcome the barrier between the low- and high-energy states of Cse1p. Consequently, the ternary complex would be spring-loaded so that cytoplasmic RanGTP hydrolysis would be sufficient to release this stored energy,

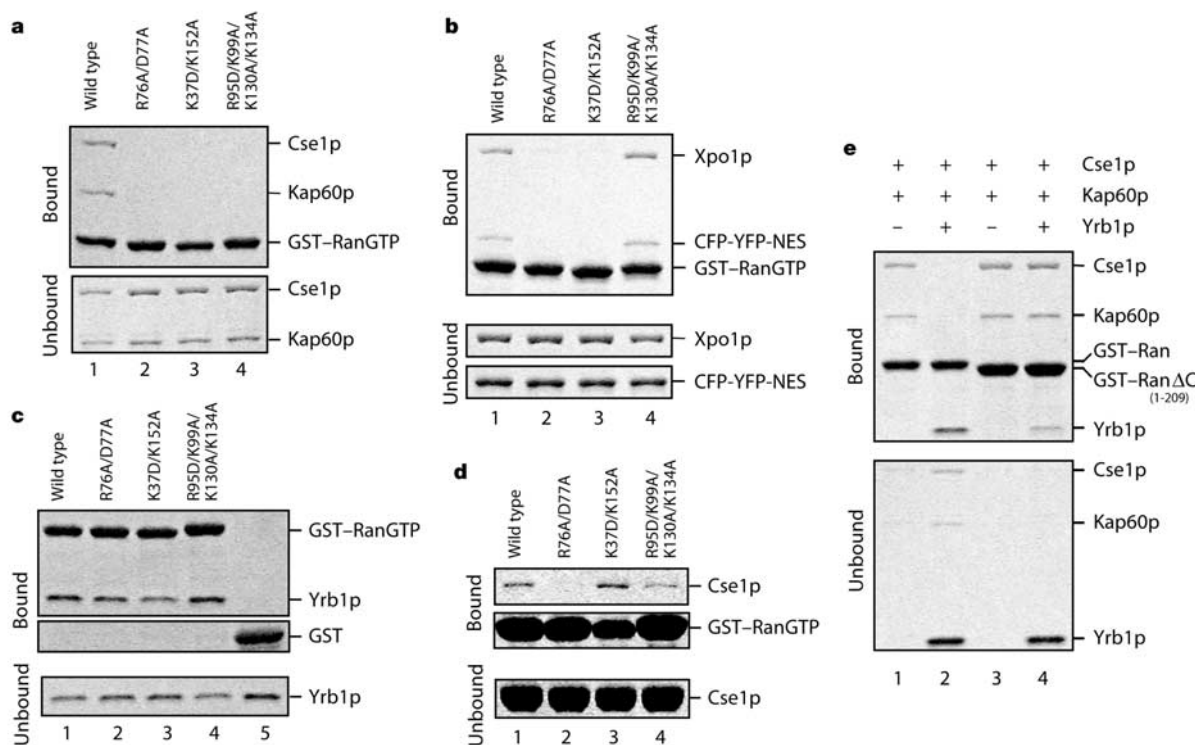


Figure 4 Mutations that influence the Cse1p:Kap60p:RanGTP complex. **a**, Ran mutations at each interface impair assembly of the ternary complex. Immobilized GST–RanGTP was treated with Cse1p and Kap60p. Whereas wild-type RanGTP bound Kap60p and Cse1p, no binding was observed with the mutants. **b**, The Ran mutation (R76A/D77A or K37D/K152A) that inhibits formation of the Cse1p:Kap60p:RanGTP complex also inhibits formation of the Xpo1p:RanGTP:NES complex. Immobilized GST–RanGTP was treated with Xpo1p and the MVM NS2 NES. Whereas wild-type RanGTP bound Xpo1p and the NES, no binding was observed with the Ran mutants. **c**, All RanGTP mutants (immobilized as GST-fusions) bound Yrb1p as wild-type (lanes 1–4) whereas GST alone (lane 5) did not. **d**, Binding of Cse1p to RanGTP in the absence of Kap60p. Immobilized GST–RanGTP was

treated with Cse1p. R76A/D77A-RanGTP inhibited Cse1p binding completely, whereas R95D/K99A/K130A/K134A-RanGTP bound to a lesser extent than wild-type or K37D/K152A-RanGTP, indicating that, in the absence of Kap60p, RanGTP did not bind to both sites on Cse1p. **e**, The Ran C terminus is critical for the disassembly of the ternary complex by Yrb1p. Immobilized GST–Ran (full-length, lanes 1 and 2) or GST–Ran Δ C (residues 1–209, lanes 3 and 4) was incubated with Cse1p:Kap60p, unbound material washed away and then treated with buffer alone (lanes 1 and 3) or Yrb1p (lanes 2 and 4). Yrb1p was able to dissociate only the complex formed with full-length RanGTP, consistent with the importance of the steric clash between the Ran C terminus.

causing it to fly apart and dissociate. Notably, both the Ran R76A/D77A or K37D/K152A mutations also inhibit formation of an Xpo1p:RanGTP:substrate complex (Fig. 4b), indicating that a two-site RanGTP clamping mechanism may be a general feature of karyopherin- β -family nuclear export factors.

In summary, the structure of the Cse1p:Kap60p:RanGTP nuclear export complex shows a remarkably intricate series of interfaces that enable molecular recognition of cargo-free Kap60p, and suggests a spring-loaded mechanism to facilitate disassembly of the complex in the cytoplasm following RanGTP hydrolysis. □

Methods

Cse1p:Kap60p:RanGTP complex

Glutathione-S-transferase (GST)—Cse1p (*S. cerevisiae* residues 1–960), His/S–Ran (canine, residues 1–176 with added His- and S- tags) and Kap60p (*S. cerevisiae* residues 1–530) were expressed separately in *Escherichia coli* strain BL21-Gold(DE3) (Stratagene) at 20 °C overnight using respectively pGEX-TEV, pET30a-TEV and pET30a vectors. pGEX-TEV has a TEV proteolysis site inserted after the thrombin site in pGEX-4T-1 (Amersham), whereas pET30a-TEV has the TEV site in place of the enterokinase site in pET30a (Novagen). Expression of Kap60p was increased by replacing some rarely used codons in the IBB domain (Arg 17, Thr 19, Arg 55 and Arg 56) with ones more frequently used in bacteria (Arg:CGT, Thr:ACC). After collection, the three sets of cells were mixed, suspended in buffer A (30 mM Tris-HCl (pH 7.5), 180 mM NaCl, 5 mM Mg(OAc)₂, 10 mM imidazole, 7 mM β -mercaptoethanol, 1 mM Pefabloc) and lysed by sonication on ice. All subsequent steps were performed at 4 °C. Clarified lysates were loaded onto Ni-NTA resin (Novagen), washed with buffer A containing 20 mM imidazole, eluted with buffer A containing 250 mM imidazole, and GTP and Tween20 added to 0.1 mM and 0.05%, respectively. After incubating the Ni-NTA eluate with Glutathione-sepharose (Amersham) overnight and washing with buffer B (10 mM Tris-HCl (pH 7.6), 100 mM NaCl, 5 mM Mg(OAc)₂, 2 mM β -mercaptoethanol, 0.05% Tween20), the GST- and His/S- tags were removed with His-TEV protease (70 μ g ml^{−1}) overnight in buffer B containing 0.1 mM GTP and 0.1 mM Pefabloc. The Cse1p:Kap60p:RanGTP complex released from the resin was passed through Ni-NTA and finally purified over Superdex200 (Amersham) in 10 mM Tris-HCl (pH 7.6), 100 mM NaCl, 5 mM Mg(OAc)₂. GTP was added to 20 μ M, and the complex concentrated to 4 mg ml^{−1} using a Millipore concentrator (*M*_w 5,000).

X-ray crystallography

Plate-shaped crystals of the Cse1p:Kap60p:RanGTP complex, typically 400 × 200 × 30 μ m and belonging to space group C2 (*a* = 272.7 Å, *b* = 72.3 Å, *c* = 97.6 Å, β = 92.6°, one complex per asymmetric unit), were obtained at 18 °C by streak seeding hanging drops containing 4 mg ml^{−1} protein, 50 mM Tris-HCl (pH 7.6), 200 mM NaCl, 5 mM Mg(OAc)₂, 100 μ M GTP, 22% PEG3350 and 2% PEG400. Freeze-thaw annealing reduced mosaic spread. A 2.0 Å data set was collected at 100 K at ESRF beamline ID23-1 (Grenoble, France) and processed using MOSFLM and CCP4 programs²¹. The structure was solved by molecular replacement using CNS²² using Kap60p (residues 89–509)²³ and Ran (residues 9–176)⁹ as models. The density map calculated using molecular replacement phases showed interpretable α -helical densities for Cse1p on the surface of Ran. Iterative cycles of model building (aided by ARP/wARP²⁴) using O²⁵ and refinement using CNS²² produced a final model with *R*_{free} 26.8% (*R*_{cryst} 23.4%, Table 1) that contained Cse1p residues 1–193, 196–248, 255–371, 376–413, 418–880 and 886–958, Kap60p residues 12–19, 31–58 and 87–513, Ran residues 7–176, MgGTP and 486 water molecules. Cse1p Glu585 was a Ramachandran outlier, but the electron density unambiguously showed the location of its carbonyl oxygen. Molecular graphics used Molscript²⁶, Raster3D²⁷ and MSMS.

Proteins for binding assays

YRBI was PCR amplified from yeast genomic DNA (Promega) and cloned into pET30a to express His/S–Yrb1p. GST–Ran (canine, full-length), GST–Cse1p (full-length) and GST–Cbp80p NLS (residues 1–30) were expressed from pGEX-TEV. GST–Ran Δ C (residues 1–209) was constructed by mutating Pro210 to a stop codon using the Quickchange (Stratagene) system. His/S–Ran (canine, full-length) was expressed from pET30a. His/S–Kap60p (full-length) was expressed as described²⁸. His/S–Xpo1p was expressed from pET30a-TEV and His-CFP-YFP-MVM NS2 NES²⁹ from pET15b. Mutants of GST–Ran and His/S–Kap60p were prepared using the Quickchange system and all constructs were verified by sequencing. Proteins were expressed in BL21-Gold(DE3) or BL21(DE3)RIL cells (Stratagene). GST-fusions were purified over Glutathione-sepharose and His-tagged proteins purified over Ni-NTA resin. Cse1p was expressed and purified as a GST-fusion and cleaved with His-TEV protease. Ran was loaded with GTP as described¹¹.

Pulldown assays

Assays were performed in PBS containing 5 mM Mg(OAc)₂, 0.1% Tween20, 2 mM DTT and 0.2 mM PMSF as described²⁸. GST-fusions were immobilized on 10 μ l of packed Glutathione-sepharose, and binding reactions were performed in a total volume of 50 μ l for 1 h at 4 °C. Beads were then centrifuged and supernatants saved (unbound fraction). Beads were washed with 2 × 1 ml of binding buffer and bound proteins eluted with SDS-sample buffer. Bound and unbound fractions were analysed by SDS–PAGE. To assay binding of Cse1p and Kap60p to RanGTP (Figs 2f and 4a), immobilized GST–RanGTP (15 μ g) was incubated with Cse1p (3 μ g) and Kap60p (3 μ g). Binding of Cse1p alone to RanGTP (Fig. 4d) was analysed using 75 μ g immobilized GST–RanGTP and 50 μ g Cse1p. For Yrb1p binding (Fig. 4c), immobilized GST–RanGTP (15 μ g) was incubated with

Yrb1p (5 μ g). To assay NLS competition (Fig. 2g), immobilized GST–Cbp80p NLS (7 μ g) was incubated with Kap60p (2 μ g), unbound Kap60p washed away and the beads then incubated with or without Cse1p (14 μ g) and RanGTP (20 μ g). To assay disassembly of the Cse1p:Kap60p:RanGTP complex by Yrb1p (Fig. 4e), immobilized GST–RanGTP (15 μ g) was incubated with Cse1p (3 μ g) and Kap60p (3 μ g), unbound material washed away, and the beads incubated with or without Yrb1p (9 μ g). For binding of Xpo1p and NES to RanGTP (Fig. 4b), immobilized GST–RanGTP (15 μ g) was incubated with Xpo1p (6 μ g) and CFP-YFP-MVM NS2 NES (6 μ g).

Received 30 September; accepted 1 November 2004; doi:10.1038/nature03144.

- Weis, K. Regulating access to the genome. Nucleocytoplasmic transport throughout the cell cycle. *Cell* **112**, 441–451 (2003).
- Mosammaparast, M. & Pemberton, L. Karyopherins: from nuclear transport mediators to nuclear function regulators. *Trends Cell Biol.* **14**, 547–556 (2004).
- Kobe, B. Autoinhibition by an internal nuclear localization signal revealed by the crystal structure of mammalian importin α . *Nature Struct. Biol.* **6**, 388–397 (1999).
- Hood, J. K. & Silver, P. A. Cse1p is required for export of Srp1/importin- α from the nucleus in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **273**, 35142–35146 (1998).
- Solsbacher, J., Maurer, P., Bischoff, F. R. & Schlenstedt, G. Cse1p is involved in export of yeast importin α from the nucleus. *Mol. Cell. Biol.* **18**, 6805–6815 (1998).
- Kutay, U., Bischoff, F. R., Kostka, S., Kraft, R. & Gorlich, D. Export of importin α from the nucleus is mediated by a specific nuclear transport factor. *Cell* **90**, 1061–1071 (1997).
- Cingolani, G., Petose, C., Weis, K. & Muller, C. Structure of importin- β bound to the IBB domain of importin- α . *Nature* **399**, 221–229 (1999).
- Chook, Y. M. & Blobel, G. Structure of the nuclear transport complex karyopherin- β 2-Ran-GppNHp. *Nature* **399**, 230–237 (1999).
- Vetter, I. R., Arndt, A., Kutay, U., Gorlich, D. & Wittinghofer, A. Structural view of the Ran-Importin β interaction at 2.3 Å resolution. *Cell* **97**, 635–646 (1999).
- Lee, S. J. *et al.* The structure of importin- β bound to SREBP-2: nuclear import of a transcription factor. *Science* **302**, 1571–1575 (2003).
- Bayliss, R., Littlewood, T. & Stewart, M. Structural basis for the interaction between FxFG nucleoporin repeats and importin- β in nuclear trafficking. *Cell* **102**, 99–108 (2000).
- Nilsson, J., Weis, K. & Kjems, J. The C-terminal extension of the small GTPase Ran is essential for defining the GDP-bound form. *J. Mol. Biol.* **318**, 583–593 (2002).
- Herold, A., Truant, R., Wiegand, H. & Cullen, B. R. Determination of the functional organization of the importin α nuclear import factor. *J. Cell Biol.* **143**, 309–318 (1998).
- Harreman, M. T. *et al.* Characterization of the auto-inhibitory sequence within the N-terminal domain of importin α . *J. Biol. Chem.* **278**, 21361–21369 (2003).
- Gruss, O. J. *et al.* Ran induces spindle assembly by reversing the inhibitory effect of importin α on TPX2 activity. *Cell* **104**, 83–93 (2001).
- Behrens, P., Brinkmann, U. & Wellmann, A. CSE1L/CAS: Its role in proliferation and apoptosis. *Apoptosis* **8**, 39–44 (2003).
- Xiao, Z., McGrew, J. T., Schroeder, A. J. & Fitzgerald-Hayes, M. CSE1 and CSE2, two new genes required for accurate mitotic chromosome segregation in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **13**, 4691–4702 (1993).
- Conti, E., Uy, M., Leighton, L., Blobel, G. & Kuriyan, J. Crystallographic analysis of the recognition of a nuclear localization signal by the nuclear import factor karyopherin α . *Cell* **94**, 193–204 (1998).
- Gorlich, D. *et al.* A novel class of RanGTP binding proteins. *J. Cell Biol.* **138**, 65–80 (1997).
- Vetter, I. R., Nowak, C., Nishimoto, T., Kuhlmann, J. & Wittinghofer, A. Structure of a Ran-binding domain complexed with Ran bound to a GTP analogue: implications for nuclear transport. *Nature* **398**, 39–46 (1999).
- Collaborative Computational Project No. 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
- Brunker, A. T. *et al.* Crystallography and NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **50**, 905–921 (1998).
- Conti, E. & Kuriyan, J. Crystallographic analysis of the specific yet versatile recognition of distinct nuclear localization signals by karyopherin α . *Struct. Fold. Des.* **8**, 329–338 (2000).
- Perrakis, A., Morris, R. & Lamzin, V. S. Automated protein model building combined with iterative structure refinement. *Nature Struct. Biol.* **6**, 458–463 (1999).
- Jones, T. A., Cowan, S. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
- Kraulis, P. J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
- Merritt, E. A. & Bacon, D. J. Raster3D: photorealistic molecular graphics. *Methods Enzymol.* **277**, 505–524 (1997).
- Matsuura, Y., Lange, A., Harreman, M. T., Corbett, A. H. & Stewart, M. Structural basis for Nup2p function in cargo release and karyopherin recycling in nuclear import. *EMBO J.* **22**, 5358–5369 (2003).
- Askjaer, P. *et al.* RanGTP-regulated interactions of CRM1 with nucleoporins and a shuttling DEAD-box helicase. *Mol. Cell. Biol.* **19**, 6276–6285 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank the staff of the European Synchrotron Radiation Facility, especially D. Nurizzo and J. McCarthy, for assistance during data collection, and our colleagues in Cambridge, especially K. Nagai, R. Henderson, R.A. Crowther, P. Evans, A. Leslie and A. Stewart for comments, criticism and assistance. This work was supported in part by a Human Frontier Science Program grant.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.S. (ms@mrc-lmb.cam.ac.uk). Atomic coordinates and structure factors have been deposited in the Protein Data Bank under accession codes 1WA5 and rlwa5sf.ent, respectively.

An extreme distortion of the Van Allen belt arising from the 'Hallowe'en' solar storm in 2003

D. N. Baker¹, S. G. Kanekal¹, X. Li¹, S. P. Monk¹, J. Goldstein² & J. L. Burch²

¹Laboratory for Atmospheric and Space Physics, University of Colorado, 1234 Innovation Drive, Boulder, Colorado 80303-7814, USA

²Southwest Research Institute, PO Drawer 28510, San Antonio, Texas 78284, USA

The Earth's radiation belts—also known as the Van Allen belts¹—contain high-energy electrons trapped on magnetic field lines^{2,3}. The centre of the outer belt is usually 20,000–25,000 km from Earth. The region between the belts is normally devoid of particles^{2–4}, and is accordingly favoured as a location for spacecraft operation because of the benign environment⁵. Here we report that the outer Van Allen belt was compressed dramatically by a solar storm known as the 'Hallowe'en storm' of 2003. From 1 to 10 November, the outer belt had its centre only ~10,000 km from Earth's equatorial surface, and the plasmasphere was similarly displaced inwards. The region between the belts became the location of high particle radiation intensity. This remarkable deformation of the entire magnetosphere implies surprisingly powerful acceleration and loss processes deep within the magnetosphere.

Sunspot group number 484 appeared on the east limb of the Sun's disk on 18 October 2003. As it and concurrent sunspot groups 486 and 488 rotated across the visible solar surface during the subsequent two to three weeks, the Sun produced spectacular enhancements of solar X-rays ('flares'), solar energetic particles, and, ultimately, some of the largest geomagnetic storms on record⁶. The interplanetary shock waves and coronal mass ejections launched by the Sun reached Earth's vicinity in time periods as short as one day, and produced dramatic effects in near-Earth space. In striking Earth's outer magnetic envelope (the magnetosphere), the high-speed solar disturbances compressed, distorted and enhanced the Van Allen belts. Figure 1a shows that there is a high degree of electron flux variability on daily, monthly and solar-cycle timescales^{7,8}. It is also clear that the maximum flux of energetic electrons does not occur at the time of highest sunspot number (2000–01), but rather in the declining phase (for example, 1994–95) of solar activity⁹. The period of October–November 2003 was a period of exceptional radiation belt variability.

Figure 1b demonstrates the remarkable degree to which the radiation belt electrons were displaced inwards towards the Earth in 2003 by the Hallowe'en storms after day of year (DOY) ~300. In fact, the data show that the normal peak of electron fluxes around $L = 4.0$ that was seen before (and well after) the Hallowe'en storm period was displaced far inward to $L \approx 2.5$ for a period of at least two weeks (see Fig. 1 legend for a definition of L). The normal slot region—the region between the belts, typically devoid of high-energy electrons^{3,4,10}—was filled with electrons for several weeks (from DOY ≈ 300 to DOY ≈ 340). Such an intense and long-lasting slot-filling event was not previously seen in the long run of SAMPEX satellite electron flux data (Fig. 1a). Note, also, from Fig. 1b that the inner zone ($1 \leq L \leq 2$) was filled with high-energy electrons to a degree not seen from 1992 to 2003, and that this has persisted to the present day (see also refs 11–13).

We find evidence that the great distortion and displacement of the outer radiation belt during and following the Hallowe'en storm was closely associated with a major reduction of the Earth's plasmasphere. The plasmasphere is the region of cold, relatively dense ionized gas (mostly protons and helium ions) that resides on the magnetic field lines close to the Earth¹⁴. It is understood that the

plasmasphere is threaded by magnetic flux tubes that are persistently 'closed', so that plasma from the Earth's ionosphere has filled the flux tubes and reached a near-equilibrium state along the entire dipole-like field line¹⁵. The outer boundary of the plasmasphere—known as the 'plasmopause'—can act to refract, and hence trap, electromagnetic waves propagating in the whistler mode¹⁶. Such confined waves in the plasmasphere strongly scatter trapped electrons into the atmospheric 'loss cone', causing the electrons to be 'precipitated' into the Earth's upper atmosphere¹⁷. Such precipitated electrons are removed from the magnetosphere permanently. It is generally believed that the radiation belt slot region (described above) is due to such strong scattering and loss of electrons near the plasmopause caused by strong wave–particle interactions^{16,17}.

The Imager for Magnetopause-to-Aurora Global Exploration (IMAGE) spacecraft¹⁸ makes it possible to view directly the physical extent of the plasmasphere by imaging the extreme ultraviolet (EUV) emissions from He⁺ (at 30.4 nm wavelength) within the plasmasphere¹⁹. With its high inclination and highly elliptical ($1,000 \times 46,000$ km altitude) orbit, IMAGE is able to view the Earth's plasma environs for extended periods¹⁸. Figure 2 (top row) shows that surrounding the Earth in each frame is a green-to-white 'cloud' that is due to the EUV light scattered by the He⁺ resident in the plasmasphere. Careful analysis^{19,20} allows the EUV images to be used to determine the outer edge of the plasmasphere—that is, the plasmopause.

The lower row of Fig. 2 shows the inferred plasmopause locations (in local time); they generally delineate very well the equatorial extent of the plasmasphere²⁰. Before the storm, on 28 October (leftmost panel), the plasmopause radius was between about 4 and 5 Earth radii ($4\text{--}5 R_E$). During the storm, on 31 October (second panel from the left), the plasmopause radius was inside $2 R_E$, and at some longitudes was at $1.5 R_E$. Such an extremely small plasmasphere only occurs during the strongest geomagnetic storms¹⁹, and the plasmopause shrinkage of the Hallowe'en 2003 storm is the most pronounced yet identified in the IMAGE data set.

Using available IMAGE data (as shown in Fig. 2), we have determined the plasmopause locations throughout late October and into mid-November 2003. The minimum L -value for this boundary in Fig. 1c shows that the plasmopause often tracks rather well with the inner extent of the electron radiation belt population measured by SAMPEX. In other words, the IMAGE/EUV data show a generally close correspondence (for the analysed period) between the inner edge of the outer Van Allen belt and the measured plasmopause location. It is widely accepted that the inner edge of the outer Van Allen belt should correspond rather closely to the plasmopause location¹⁶. Our data for the (post) Hallowe'en storm period show this to very much be true. Figure 1d emphasizes the point that the electron fluxes in the normal slot region ($L \approx 2.5$) were considerably higher in early November 2003 than were the fluxes in the normal 'heart' of the outer zone ($L \approx 4.0$).

Figure 3a shows schematically the radiation belt and plasmasphere under typical conditions. Under these normal circumstances, the outer radiation belt extends quite far from the Earth with its maximum intensity occurring at $\sim 4\text{--}5 R_E$ geocentric distance. The plasmasphere extends outward to $3\text{--}4 R_E$ geocentric distance. As shown in Fig. 3b, we find during (and following) the Hallowe'en 2003 storm that the plasmasphere was greatly contracted inward towards the Earth and remained in a very reduced state for many days. Concurrently, the Van Allen belt was also re-established very far inwards and the highest electron intensities were seen at $2\text{--}3 R_E$ geocentric distance.

An event of such magnitude and extent as the October–November 2003 storms provides a unique opportunity to study the acceleration, transport and loss of energetic particles in Earth's magnetosphere. In many ways, the whole state of the Earth's radiation belts was altered in a very short time by the Hallowe'en

storms. On the other hand, the dramatically altered particle fluxes during the storm had very long-lasting consequences.

An important point to note (as shown in Fig. 1) is that the immediate consequence of solar effluents hitting Earth's magnetosphere (on ~31 October 2003) was to deplete the radiation belts at almost all L -values. Then, in the subsequent day or two the radiation belt electron fluxes were regenerated to remarkably high levels. However, this regenerated population of electrons was formed in a wholly unexpected location—the normal slot region, where few electrons can ordinarily survive for any extended period. The powerful acceleration of electrons that had to occur over DOY ≈ 305 –315 was associated with very strong wave activity measured

both in space by the Polar spacecraft²¹ and on the Earth's surface by ground-based magnetometers (I. Mann, personal communication). Thus, we would expect that waves in the ultra-low frequency (ULF) range could drive rapid radial diffusion of electrons (hence accelerating the particles)²². However, it is also possible that strong 'chorus' emissions (in the whistler mode) or even electromagnetic ion cyclotron (EMIC) waves could locally heat and accelerate electrons to very high energies^{23,24} in the $L = 2$ –3 range under these exceptional conditions.

We emphasize that losses of the radiation belt electrons can be well studied by such a dramatic and sharply defined set of events as the Hallowe'en storm. As we have shown here, the inner extent of the 'new' radiation belt electron population at $2 \leq L \leq 3$ corresponds quite closely to the outward extent of the plasmasphere. Thus, even under the conditions of extreme distortion of the inner magnetosphere, it seems that strong local acceleration plus wave-particle scattering and loss near the plasmapause must determine where the outer belt electrons can persist^{16,17}.

On the basis of the exponential decay of electron fluxes seen in November 2003 at $L \approx 2.5$ (as shown here in Fig. 1d), we can estimate the 'e-folding' lifetime of 2–6 MeV electrons from the following equation: $J(E = 2\text{--}6 \text{ MeV}) = K \exp(-t/t_0)$ where J is electron flux, K is a proportionality constant, and t_0 is the 'lifetime' of the electrons against precipitation loss (or radial diffusion away from $L = 2.5$). From a simple fitting procedure, we find that from 3 to 20 November, $t_0 \approx 4.6$ days. Following the other major storm enhancement that occurred on 20 November 2003, there was another decay over the interval 25 November to 20 December 2003. This led to $t_0 \approx 2.9$ days. Thus, the 'active' acceleration and decay episodes produced within the inner magnetosphere by the Sun's output gives us a direct measure of the loss lifetimes of electrons in the inner magnetosphere¹⁶.

A final point to note about losses is that many electrons obviously passed through the normal slot-region 'barrier' following the Hallowe'en storm and entered the inner zone ($1 \leq L \leq 2$) region (see Fig. 1b, c). These electrons constitute a new, powerful, population of very energetic electrons in the inner zone that have not been present there since the remnants of the March 1991 storm died away^{11–13}.

The presence of very energetic electrons in the Earth's magnetosphere constitutes a 'space weather' hazard to spacecraft operating in near-Earth space⁵. The Hallowe'en storms generated large fluxes

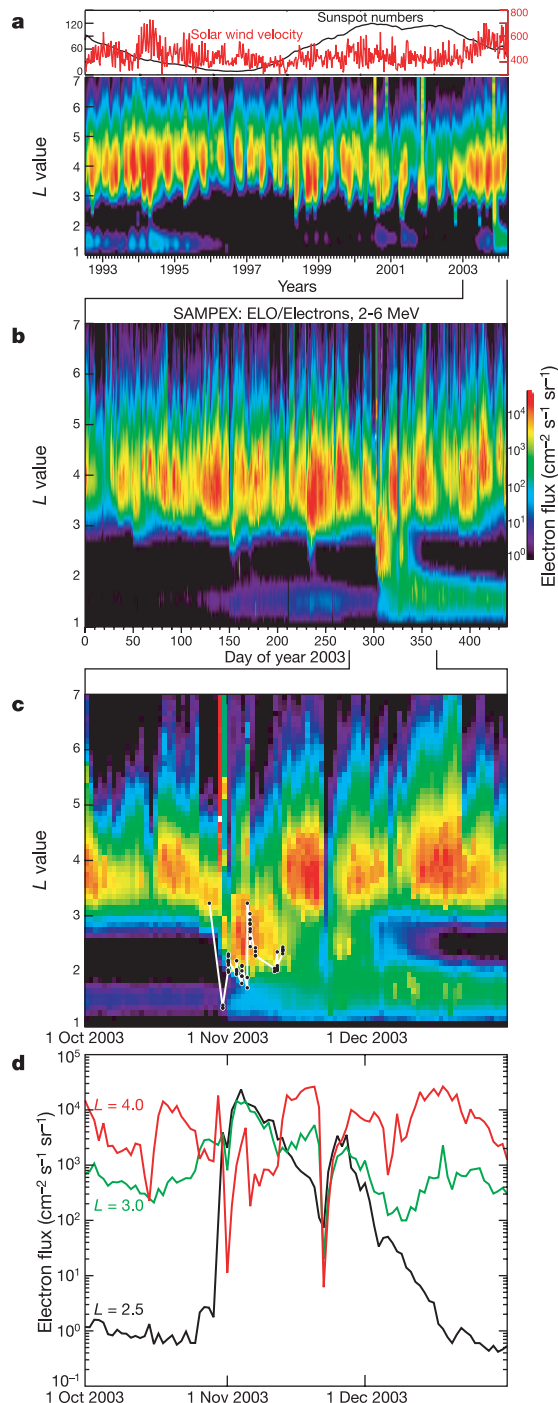


Figure 1 Energetic electron data from a low-altitude Earth-orbiting spacecraft showing both a long historical record of the Van Allen radiation belts and the specific effects of the 2003 Hallowe'en storm. **a**, Flux of 2–6 MeV electrons measured by the Solar, Anomalous, and Magnetospheric Particle Explorer (SAMPEX)⁷ from 6 July 1992 to March 2004. Data are coded according to the colour bar to the right. Vertical axis is the L -shell parameter, which is effectively the distance in Earth radii at which a magnetic field line crosses the magnetic equatorial plane. The small upper panel shows (in black) the yearly-averaged solar sunspot number (SSN) and (in red) the weekly-averaged solar wind speed (V_{sw}) measured upstream of the Earth⁸. This panel demonstrates intense electron acceleration events (due to high-speed solar wind) in 1993–95 for $3 \leq L \leq 5$. During sunspot minimum (1996), there were significant electron events only briefly around the spring and autumn equinoxes⁹. A weak, but persistent, feature in the inner zone died away throughout 1992 and 1993; it intensified in 1994. This may, in part, be related to the 'new' radiation belt created during the March 1991 storm^{11,12}. **b**, Detailed SAMPEX electron data (daily-averaged) for 2003 and early 2004 showing the shifted position of the outer Van Allen zone and the filling of the slot region. **c**, SAMPEX data for the period 1 October to 31 December 2003. The broken line superimposed on the SAMPEX data shows the plasmapause as measured by IMAGE. **d**, Electron fluxes in the energy range $2 \leq E \leq 6$ MeV for three L -values (red: $L = 4.0$; green: $L = 3.0$; and black: $L = 2.5$) for the period 1 October to 31 December 2003. The energetic electron intensities at

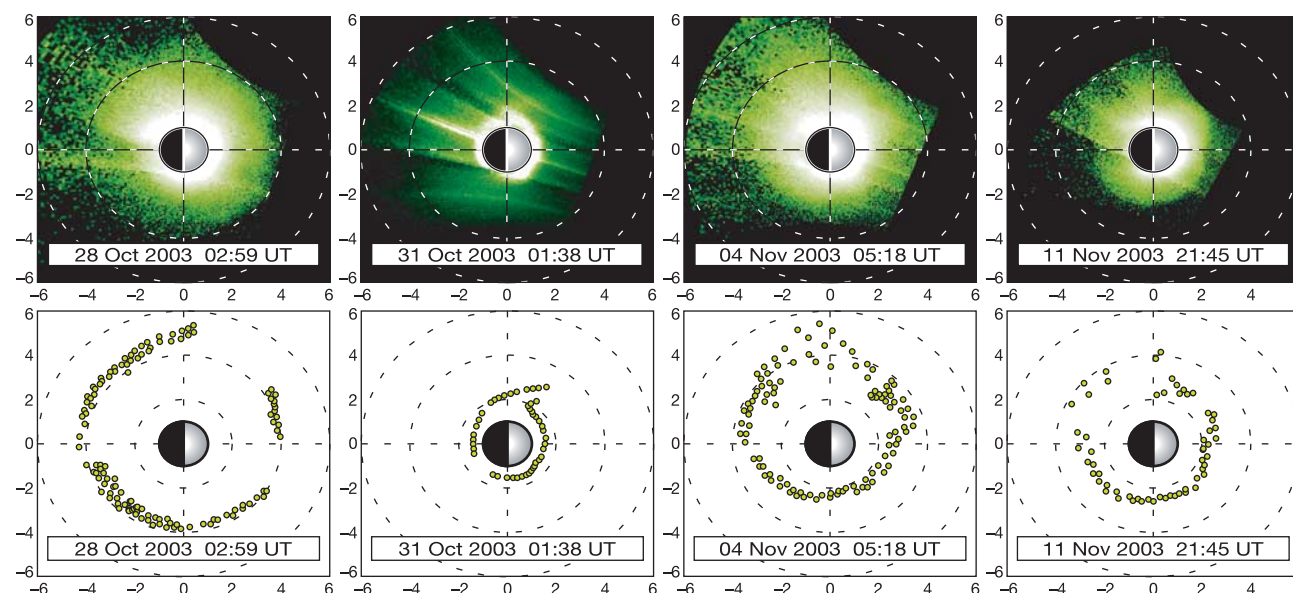


Figure 2 Extreme ultraviolet (EUV) images of the Earth's plasmasphere during late October and early November 2003. Images in the top row range in time from before the Halloween storm (28 October 2003) to well after the main storm period (11 November 2003). The Earth position is illustrated in each frame, showing the sunward and night sides by each central shaded half-circle. The effects of the Halloween storm are clearly evident in the series of plasmasphere snapshots. Another effect of the storm is in the quality of the images obtained by the EUV instrument. During the storm, an abundance of energetic particles led to radiation penetrating into the EUV cameras, creating noise in the

images. In the 31 October plasmasphere image, curved diagonal streaks of noise due to this effect are evident, cutting across the image from top-left to bottom-right. In the several days after the storm (4 November and 11 November, right two panels), the plasmasphere recovered from its extreme shrinkage, as plasma from the Earth's ionosphere gradually leaked out into space and repopulated the plasmaspheric flux tubes. The lower row of panels shows the plasmopause locations inferred at each time from each corresponding image on the top. The plasmopause positions are plotted in the equatorial x - y plane where the scale in each frame is in terms of Earth radii ($1 R_E = 6,372$ km).

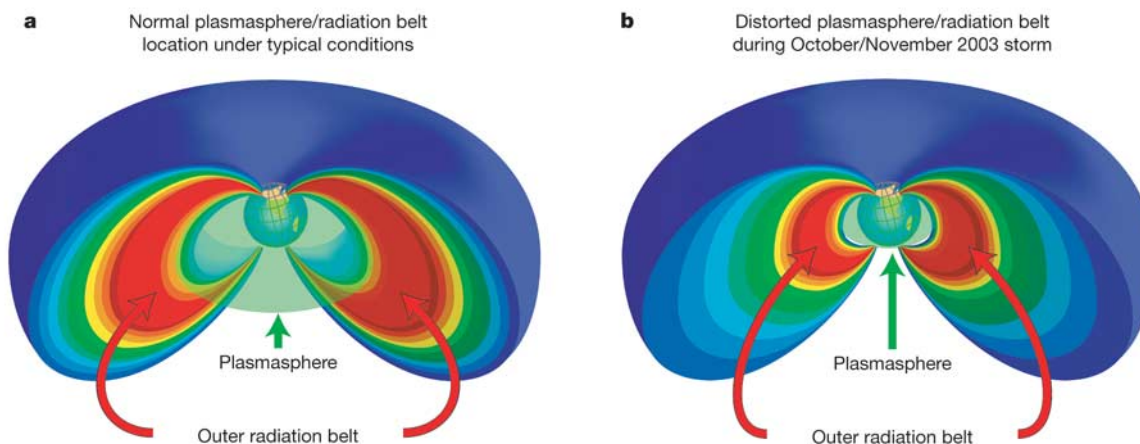


Figure 3 Diagrams of the three-dimensional view of the Earth's outer radiation belt and its relationship to the plasmasphere. **a**, A schematic diagram showing the Earth, the outer radiation belt, and the normal plasmaspheric location. **b**, Similar to **a** but showing the highly contracted plasmasphere and greatly distorted radiation belt location during the

Halloween storm period. These cutaway diagrams show the highest electron fluxes as red and the lowest fluxes as blue. The radiation belt is a 'doughnut' or torus-shaped entity. The Earth is portrayed at the centre. Also shown, as a translucent green torus towards the centre in each part of the figure, is the plasmasphere.

of solar energetic particles as well as the high fluxes of radiation belt electrons reported here. The combination of solar and magnetospheric particles produced a whole host of 'anomalies' in technological systems during October–November 2003²⁵. In fact, a few outright spacecraft failures and other significant losses were directly attributable to this space weather. We have shown that a set of events (as occurred with the Halloween storm) substantially changed the radiation environment at $L \leq 3$, and very hostile space weather suddenly developed in a region that is normally quiescent. So future models and design strategies need to take account of possible extreme events, such as occurred in October–November 2003. □

Received 6 July; accepted 12 October 2004; doi:10.1038/nature03116.

1. Van Allen, J. A. in *Discovery of the Magnetosphere* (eds Gillmor, C. S. & Spreiter, J. R.) 235–251 (Vol. 7, History of Geophysics, American Geophysical Union, Washington DC, 1997).
2. Vette, J. I. *The AE-8 Trapped Electron Model Environment* (NSSDC WDC-A-R&S 91-24, NASA Goddard Space Flight Center, Greenbelt, Maryland, 1991).
3. Brautigam, D. H., Gussenhoven, M. S. & Mullen, E. G. Quasi-static model of outer zone electrons. *IEEE Trans. Nucl. Sci.* **39**, 1797–1803 (1992).
4. Lyons, L. R. & Thorne, R. M. Equilibrium structure of radiation belt electrons. *J. Geophys. Res.* **78** (13), 2142–2149 (1973).
5. Baker, D. N. How to cope with space weather. *Science* **297**, 1486–1487 (2002).
6. Lopez, R. E., Baker, D. N. & Allen, J. H. Sun unleashes Halloween storm. *Eos* **85**, 105, 108 (2004).
7. Baker, D. N. *et al.* An overview of the Solar Anomalous, and Magnetospheric Particle Explorer (SAMPEX) mission. *IEEE Trans. Geosci. Remote Sens.* **31**, 531–541 (1993).

8. Li, X. *et al.* The predictability of the magnetosphere and space weather. *Eos* **84**, 361, 369–370 (2003).
9. Baker, D. N., Kanekal, S. G., Pulkkinen, T. I. & Blake, J. B. Equinoctial and solstitial averages of magnetospheric relativistic electrons: A strong semiannual modulation. *Geophys. Res. Lett.* **26**, 3193–3196 (1999).
10. Blake, J. B. *et al.* New high temporal and spatial resolution measurements by SAMPEX of the precipitation of relativistic electrons. *Adv. Space Res.* **18**, 171–177 (1996).
11. Looper, M. D. *et al.* Observations of remnants of the ultrarelativistic electrons injected by the strong SSC of 24 March 1991. *Geophys. Res. Lett.* **21**, 2079–2082 (1994).
12. Li, X. *et al.* Simulation of the prompt energization and transport of radiation particles during the March 23, 1991 SSC. *Geophys. Res. Lett.* **20**, 2423–2426 (1993).
13. Li, X. in *Proc. 6th Int. Conf. on Substorms (ICS6)* (ed. Winglee, R. M.) 305–311 (Univ. Washington Press, Seattle, Washington, 2000).
14. Carpenter, D. L. & Park, C. G. What ionospheric workers should know about the plasmapause/plasmasphere. *Rev. Geophys.* **11**, 133–154 (1973).
15. Grebowsky, J. M. Model study of plasmapause motion. *J. Geophys. Res.* **75**, 4329–4333 (1970).
16. Thorne, R. M., Smith, E. J., Burton, R. K. & Holzer, R. E. Plasmaspheric hiss. *J. Geophys. Res.* **78** (10), 1581–1598 (1973).
17. Lorentzen, K. R., Blake, J. B., Inan, U. S. & Bortnik, J. Observations of relativistic electron microbursts in association with VLF chorus. *J. Geophys. Res.* **A 106**, 6017–6027 (2001).
18. Burch, J. L. *et al.* Views of the Earth's magnetosphere with the IMAGE satellite. *Science* **291**, 619–624 (2001).
19. Sandel, B. R., Goldstein, J., Gallagher, D. L. & Spasojevic, M. Extreme ultraviolet imager observations of the structure and dynamics of the plasmasphere. *Space Sci. Rev.* **109**, 25–46 (2003).
20. Goldstein, J. *et al.* Simultaneous remote sensing and *in situ* observations of plasmaspheric drainage plumes. *J. Geophys. Res.* **109**, doi:10.1029/2003JA010281 (2004).
21. Russell, C. T., Cartwright, M. & Kanekal, S. The formation of the inner zone trapped radiation belt. *Adv. Space Res.* (in the press).
22. Liu, W. W., Rostoker, G. & Baker, D. N. Internal acceleration of relativistic electrons by large-amplitude ULF pulsations. *J. Geophys. Res.* **104**, 17391–17407 (1999).
23. Summers, D., Thorne, R. M. & Xiao, F. Relativistic theory of wave-particle resonant diffusion with application to electron acceleration in the magnetosphere. *J. Geophys. Res.* **A 103**, 20487–20500 (1998).
24. Summers, D. & Thorne, R. M. Relativistic electron pitch-angle scattering by electromagnetic ion cyclotron waves during geomagnetic storms. *J. Geophys. Res.* **108**, doi:10.1029/2002JA009489 (2003).
25. Webb, D. F. & Allen, J. H. Spacecraft and ground anomalies related to the October–November 2003 solar activity. *Space Weather* **2**, doi:10.1029/2004SW000075 (2004).

Acknowledgements This work was supported by NASA. We thank members of the SAMPEX and IMAGE science teams for data and discussions.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.N.B. (daniel.baker@lasp.colorado.edu).

Hall-effect evolution across a heavy-fermion quantum critical point

S. Paschen¹, T. Lühmann¹, S. Wirth¹, P. Gegenwart¹, O. Trovarelli¹, C. Geibel¹, F. Steglich¹, P. Coleman² & Q. Si³

¹Max Planck Institute for Chemical Physics of Solids, Nöthnitzer Straße 40, D-01187 Dresden, Germany

²Center for Materials Theory, Department of Physics and Astronomy, Rutgers University, Piscataway, New Jersey 08855, USA

³Department of Physics and Astronomy, Rice University, Houston, Texas 77005-1892, USA

A quantum critical point (QCP) develops in a material at absolute zero when a new form of order smoothly emerges in its ground state. QCPs are of great current interest because of their singular ability to influence the finite temperature properties of materials. Recently, heavy-fermion metals have played a key role in the study of antiferromagnetic QCPs. To accommodate the heavy electrons, the Fermi surface of the heavy-fermion paramagnet is larger than that of an antiferromagnet^{1–3}. An important unsolved question is whether the Fermi surface transformation at the QCP develops gradually, as expected if the magnetism is of spin-density-wave (SDW) type^{4,5}, or suddenly, as expected if the heavy electrons are abruptly localized by magnetism^{6–8}. Here we report measurements of the low-temperature Hall coefficient

(R_H)—a measure of the Fermi surface volume—in the heavy-fermion metal YbRh_2Si_2 upon field-tuning it from an antiferromagnetic to a paramagnetic state. R_H undergoes an increasingly rapid change near the QCP as the temperature is lowered, extrapolating to a sudden jump in the zero temperature limit. We interpret these results in terms of a collapse of the large Fermi surface and of the heavy-fermion state itself precisely at the QCP.

The compound YbRh_2Si_2 investigated here appears to be one of the best suited heavy-fermion metals known in which to study the evolution of the Hall effect across a QCP. Magnetic susceptibility and specific heat indicate that this compound orders antiferromagnetically via a second-order phase transition at very low temperatures (Néel temperature $T_N = 70$ mK)⁹. The antiferromagnetic nature of the transition is supported by NMR data¹⁰. Neutron scattering experiments to directly detect the magnetic order are not available to date, presumably owing to the smallness of the ordered moment¹¹. T_N is continuously suppressed down to the lowest experimentally accessed temperatures by application of a small magnetic field B_c ($B_{1c} \approx 0.7$ T for a field along the magnetically hard c axis, $B_{2c} \approx 60$ mT for a field within the easy tetragonal plane)¹². In addition, isothermal magnetostriction measurements indicate that the transition remains of second order down to at least 15 mK (ref. 13). Although a change from second to first order at even lower temperatures can, of course, not be strictly ruled out, the non-Fermi liquid behaviour observed for three decades of temperature ($10 \text{ mK} < T < 10 \text{ K}$)¹² is best described within a quantum critical picture. The use of tiny fields permits one to reversibly access the QCP without the introduction of additional disorder and without altering the character of the underlying zero-field transition¹⁴. Moreover, unlike the case of several other heavy-fermion compounds¹⁵ (and the high- T_c superconductors), the QCP is not hidden by superconductivity. This is in spite of the high quality of the YbRh_2Si_2 single crystals investigated here (residual resistivities of $\sim 1 \mu\Omega \text{ cm}$, ref. 12). The scaling analysis of the thermodynamic and dynamical properties (specific heat, magnetic susceptibility, electrical resistivity) suggests¹² that the field-induced QCP in YbRh_2Si_2 is of local^{6–8} rather than of itinerant^{4,5} type, similar to the doping-induced QCP in $\text{CeCu}_{6-x}\text{Au}_x$ (ref. 6). Hall-effect measurements may be used to access a static electronic property, namely the Fermi surface volume, for which clear-cut theoretical predictions exist for different types of QCP^{7,8,16}. The study presented here is, to our knowledge, the first systematic Hall-effect measurement at a heavy-fermion QCP.

The Hall effect, usually a rather complex quantity, appears to be surprisingly simple here, both in vanishingly small and in finite magnetic fields. Outside the quantum critical region, the Hall resistivity is linear in field, resembling the behaviour of simple metals. Furthermore, our analysis of the temperature-dependent Hall coefficient in terms of the anomalous Hall effect (Fig. 1a, Methods, and refs 17 and 18) reveals that the low-temperature (below about 1 K) Hall coefficient is dominated by its normal contribution. These features imply that the low-temperature Hall coefficient can be used, to a good approximation, as a measure of the Fermi surface volume. In the absence of photoemission and de Haas–van Alphen studies (the latter presumably never being available because of the very low critical magnetic field), and of electronic bandstructure calculations, this is, so far, the only information on the Fermi surface volume of YbRh_2Si_2 . At zero magnetic field, the data measured at the lowest temperatures tend to saturate at the value of the normal Hall coefficient extracted from the data between 7 K and room temperature (Fig. 1a). This indicates that, at $B = 0$, the Fermi surface volume is the same at the lowest temperatures as it is at high temperatures. Thus, even though there is evidence for the onset of Kondo screening at approximately 20 K (refs 9 and 12) and for surprisingly large effective quasiparticle masses in the antiferromagnetically ordered state close to the QCP¹², the local moments do not, at the lowest temperatures and at $B = 0$

appear to be incorporated into the Fermi surface. In the static sense^{1–3}, YbRh₂Si₂ may, in its unconventional antiferromagnetic state at $B < B_0$, therefore not be classified as a ‘heavy-fermion’ metal.

In the intermediate temperature range, between approximately 70 mK and 7 K, there is an additional contribution ΔR_H which is not due to the anomalous Hall effect (Fig. 1a). In the main part of Fig. 1b we show that, between 0.7 and 5 K, the cotangent of the Hall

angle is linear in T^2 (while $\Delta\rho \propto T$), indicating that the longitudinal and transverse scattering rates are different¹⁹. This type of behaviour is well known in high- T_c copper oxides, where it has been taken as evidence for spin-charge separation¹⁹. Note, however, that for YbRh₂Si₂ the temperature range where this relation holds (inset of Fig. 1b), is narrower than that where the non-Fermi liquid (NFL) behaviour $\Delta\rho \propto T$ is observed (from 100 mK to almost 10 K, ref. 12), even if YbRh₂Si₂ is field-tuned to quantum criticality

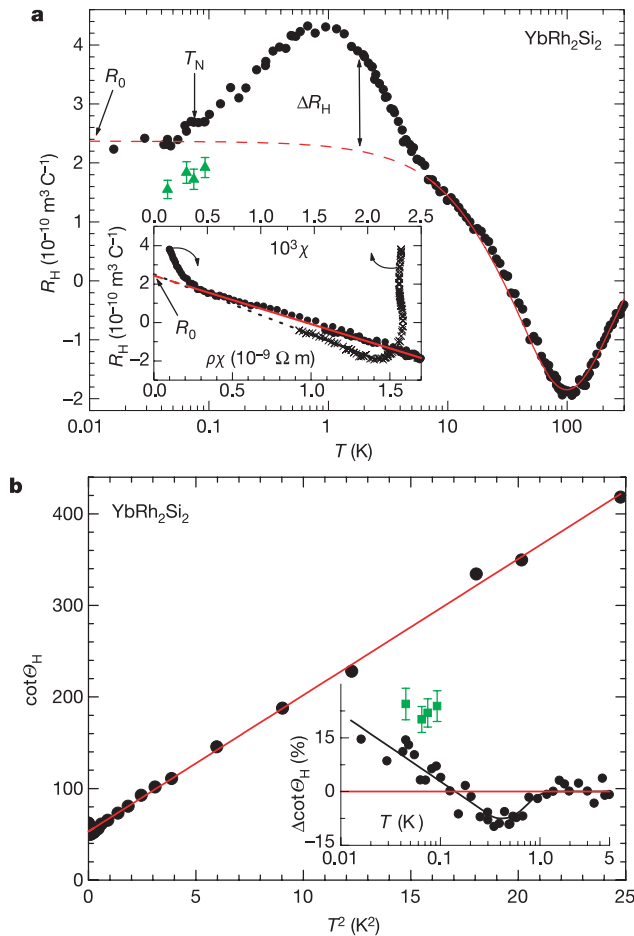


Figure 1 Temperature dependence of the Hall effect of YbRh₂Si₂. **a**, Temperature-dependent initial Hall coefficient $R_H(T)$, obtained from the initial slope of Hall resistivity versus field isotherms (Fig. 2a). The red curve corresponds to the red fit to the data from the inset. ΔR_H designates the difference between the data and the extrapolation of the fit. The green triangles correspond to R_H data obtained from the crossed-field experiment for large values of the tuning field B_2 (R_H^∞ values of fits to $R_H(B_2)$, see text and Fig. 2b), suggesting that the Fermi surface volume is distinctly larger in the field-induced paramagnetic state than in the antiferromagnetic state. Inset in **a**, initial Hall coefficient R_H versus product of electrical resistivity ρ and magnetic susceptibility χ (lower axis) and versus χ (upper axis), where temperature is an internal parameter. The full red (black) line is a linear fit according to the anomalous Hall-effect relation equation (3) (equation (5)) to the data between 7 and 300 K (90 and 300 K), the dashed lines are the extrapolations to $T = 0$. **b**, Cotangent of the Hall angle $\cot \theta_H (= \rho/(R_H B))$ as a function of T^2 , taken at $B = 1$ T. The red line (also in the inset) corresponds to a fit, $\cot \theta_H = C_1 + C_2 T^2$, where C_1 and C_2 are constants. Inset in **b**, difference between data and fit (red line) of main panel. The black line is a guide to the eye. Below 0.7 K, the data deviate considerably from the fit. The green squares correspond to $\cot \theta_H$ data obtained from the crossed-field experiment at the respective crossover fields ($B_2 = B_0$), indicating that, closer to the QCP, these deviations are even stronger. Thus, the $\cot \theta_H = C_1 + C_2 T^2$ behaviour appears to be a property of the regime at elevated temperatures where quantum critical fluctuations start to influence the physical properties, but it does not extend over the entire temperature region down to the QCP. Error bars, standard errors.

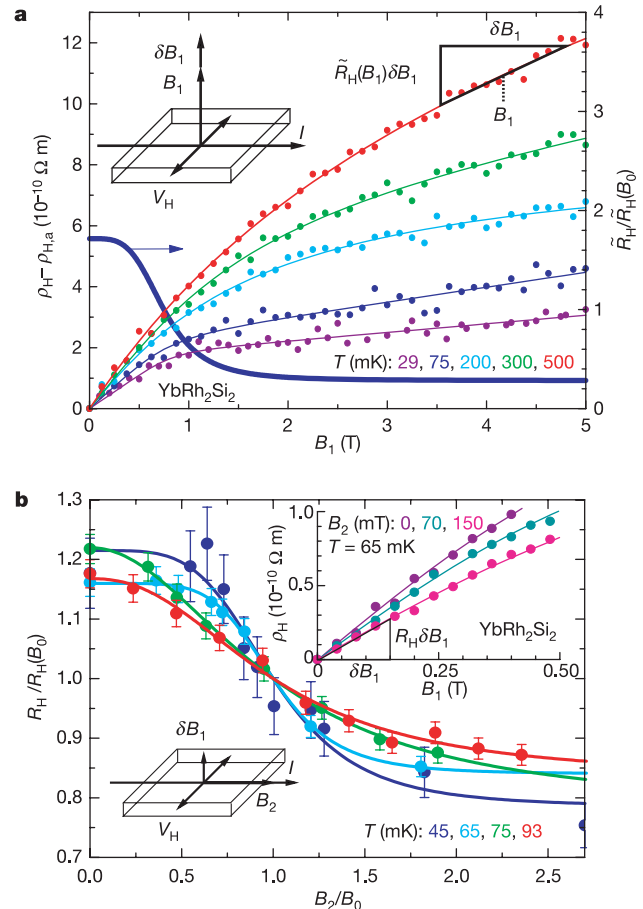


Figure 2 Magnetic field dependence of the Hall effect of YbRh₂Si₂. **a**, Single-field experiment. Typical isotherms of the Hall resistivity ρ_H , corrected for its anomalous contribution $\rho_{H,a}(B)$ (equation (6)), versus magnetic field $B_1 = \mu_0 H_1$ ($\parallel c$ -axis). The solid curves represent best fits, $\int \tilde{R}_H(B) dB$ (see text), to the data. The derivative of the fit at 75 mK is plotted on the right axis. **b**, Crossed-field experiment. Initial slope R_H , normalized to its value at the crossover field B_0 , of all measured ρ_H versus B_1 curves as a function of B_2/B_0 , at 45, 65, 75 and 93 mK. The solid lines represent best fits (see text) to the data. R_H decreases by a factor of ~ 1.5 upon going from the zero-field antiferromagnetic state to the field-induced paramagnetic state. In a spin-density-wave (SDW) picture, R_H is expected to evolve as the magnetic order parameter⁷. In YbRh₂Si₂, the ordered moment at $B = 0$ was estimated to be $\sim 0.002 \mu_B$ per Yb (ref. 11). Thus, the change in R_H corresponds to a factor of $\sim 750 \mu_B^{-1}$. The corresponding change of R_H by a factor of $\sim 30 \mu_B^{-1}$ observed for the SDW system Cr_{1-x}V_x (refs 22 and 26) was already considered a giant effect, possibly connected with bandstructure nesting effects³⁰. By comparison, the effect in YbRh₂Si₂ is about 25 times as large. Even in the absence of both experimental and theoretical studies of the electronic bandstructure of YbRh₂Si₂, we judge this effect far too large to be accounted for within an SDW picture. The same argument holds even if the second-order transition observed in the measured temperature range (> 15 mK) turned over into a first-order one at $T < 15$ mK. The inset in **b** displays ρ_H versus B_1 curves at three different values of the tuning field $B_2 = \mu_0 H_2$ ($\perp c$ -axis) at 65 mK. The solid lines represent best fits, as in **a**. Similar data have been obtained at the other temperatures (not shown). The sketches in **a** and **b** illustrate the experimental set-up. Error bars, standard errors.

(green squares in inset). The same may hold true for CeCoIn₅ (ref. 20).

In our field-dependent Hall-effect measurements on YbRh₂Si₂, the magnetic field plays dual roles, as both a ‘tuning’ and a ‘probe’ field. On the one hand, the coupling between the field and the Yb³⁺ moments tends to align the latter: it is this Zeeman-like coupling that tunes the ground state of the material, ultimately suppressing the antiferromagnetism and creating the QCP. On the other hand, the magnetic field also generates a weak Lorentz force on the underlying electrons, which produces the Hall response. The weak orbital coupling responsible for the Lorentz force does not appreciably change the ground state, so that, to a good approximation, we can discuss the two couplings independently. The single crystals of YbRh₂Si₂ are thin platelets oriented along the *a*–*b* plane, and practical Hall-effect measurements require a current and Hall voltage lying in this plane. This allows for two distinct types of experiment, namely ‘transverse tuning’ where the tuning field *B*₁ is parallel to the *c* axis, perpendicular to the current, and ‘longitudinal tuning’ where the tuning field *B*₂ lies parallel to the current in the basal plane (compare schematics in Fig. 2a and b). The longitudinal field *B*₂ produces essentially no Hall response (see Supplementary Methods 1), and serves only to tune the state: a separate, crossed probe field δB_1 along the *c* axis is required to measure the Hall response. In this longitudinal (crossed-field) experiment, the Hall resistivity ρ_H is a direct measure of the field-tuned (linear-response) Hall coefficient $R_H(B_2)$:

$$R_H(B_2) \equiv \lim_{B_1 \rightarrow 0} \rho_H(B_2, B_1) / B_1 \quad (1)$$

In the transverse (single-field) case, on the other hand, the

magnetic field simultaneously tunes the state and probes the Hall response, and the differential Hall coefficient $\tilde{R}_H(B_1)$ is:

$$\begin{aligned} \tilde{R}_H(B_1) &\equiv \frac{d\rho_H(B_1)}{dB_1} = \left[\frac{\partial \rho_H(B_1)}{\partial B_1} \right]_{\text{orb}} + \left[\frac{\partial \rho_H(B_1)}{\partial B_1} \right]_{\text{zeeman}} \\ &= R_H(B_1) + \left[\frac{\partial \rho_H(B_1)}{\partial B_1} \right]_{\text{zeeman}} \end{aligned} \quad (2)$$

The orbital (‘probing’) contribution is, according to the Kubo formalism, just the generalized definition of a Hall coefficient (see Supplementary Methods 2). The Zeeman (‘tuning’) term is not related to a readily measurable linear-response quantity.

We first discuss the results of the single-field experiment. Figure 2a displays several representative isotherms of the Hall resistivity ρ_H , corrected for its anomalous contribution $\rho_{H,a}(B)$ (see Methods), versus *B*₁. $\rho_H - \rho_{H,a}$ shows a linear low-*B*₁ behaviour with larger slope, and a linear high-*B*₁ behaviour with smaller slope. The crossover between the two regimes broadens and shifts to higher *B*₁ with increasing temperature. For a quantitative analysis of the data we choose $\tilde{R}_H(B) = R_H^\infty - (R_H^\infty - R_H^0)\gamma(B)$ as a fitting function, where R_H^0 is the zero-field Hall coefficient and R_H^∞ is the asymptotic differential Hall coefficient at large fields. $\gamma(B)$ is a crossover function that changes from unity at low fields to zero at large fields, which we parameterize as $\gamma(B) = 1/[1 + (B/B_0)^p]$. Here, *B*₀ is the crossover field and *p* determines the sharpness of the transition, which has a width $\Gamma \approx B_0/p$ when *p* is large. For $p \rightarrow \infty$, $\int \tilde{R}_H(B)dB$ has a sharp kink at *B* = *B*₀, corresponding to a step in

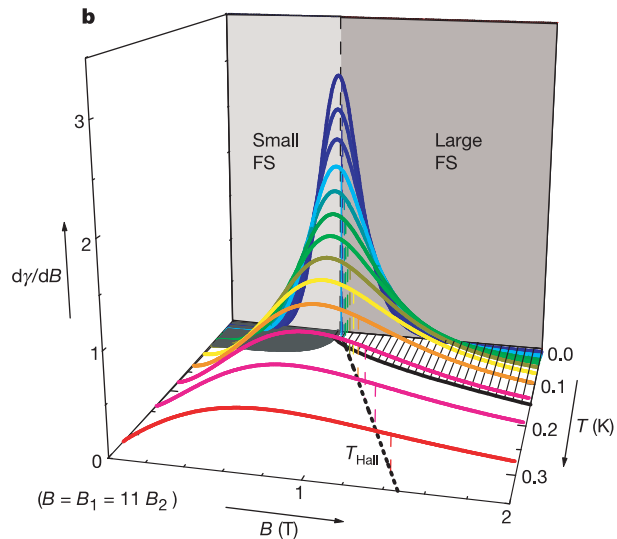
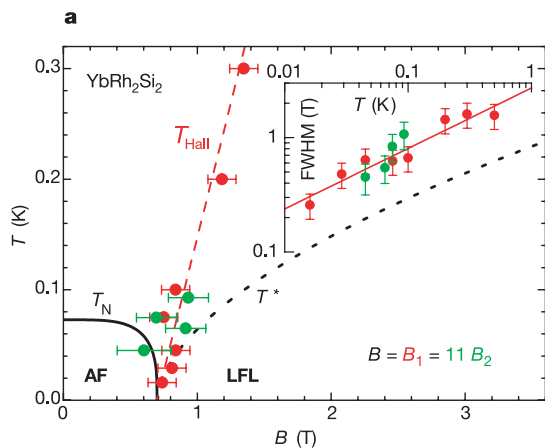


Figure 3 Temperature–field phase diagrams of YbRh₂Si₂. **a**, The red data points correspond to the *B*₀ values (crossover positions in the Hall-effect measurements) determined from the fits to the data in Fig. 2a (single-field experiment). Note that the horizontal bars represent the error in the determination of *B*₀ rather than the width of the crossover. The red dotted line denoted *T*_{Hall} is the best linear fit to all data up to 0.5 K. It extrapolates at zero temperature to ~ 0.7 T, the critical field *B*_{1c} for the direction parallel to the *c*-axis. The green data points correspond to 11*B*₀ determined from the fits to the data in Fig. 2b (crossed-field experiment). The full and dotted black curves represent the field dependence of the Néel temperature *T*_N and the crossover temperature *T*^{*} to a $\Delta\rho \propto T^2$ law, respectively, as determined from iso-field $\rho(T)$ data¹². The latter differs qualitatively from the cross-over line determined from a scaling analysis of both specific heat and resistivity data, yielding $T_{\text{cross}} \propto (B - B_c)$ (ref. 14). The inset shows the full-width at half-maximum (FWHM) of $d\tilde{R}_H(B_1)/dB_1$ in a log–log plot (red points). The red

solid line, $\propto T^a$, $a = 0.5 \pm 0.1$, is a best fit to these data. As in the main panel, the green dots correspond to the crossed-field experiment. For both the main panel and the inset, the red and green data points agree within the error bars. **b**, Three-dimensional representation of the field derivative of the crossover function $\gamma(B)$ defined in the text. The coloured curves represent arbitrary isotherms of $d\gamma(B)/dB$, obtained using both the *B*₀(*T*) fit of **a** and a power law fit to the corresponding $\rho(T)$ data (not shown). The field *B* corresponds to *B*₁||*c* or to 11*B*₂ ⊥ *c*. The positions *B*₀ are designated by broken drop lines and the black dotted line denoted *T*_{Hall} in the *T*–*B* plane. The antiferromagnetic (AF) phase and the region where $\Delta\rho \propto T^2$ (LFL) are marked as black and hatched areas, respectively, in the *T*–*B* plane. At the lowest temperatures, $d\gamma(B)/dB$ may be interpreted as indicating the change of the effective carrier concentration. In the limit $T \rightarrow 0$, $d\gamma(B)/dB$ is a δ -function (dotted line in the *T* = 0 plane), separating the states of small and large Fermi surface (FS) at *B* = *B*_{1c} = 11*B*_{2c}. Error bars, standard errors.

$\tilde{R}_H(B)$ itself. The fits to the data are shown as solid lines in Fig. 2a. For one temperature the derivative of the fit, corresponding to $\tilde{R}_H(B_1)$, is shown as well. The crossover fields B_0 obtained from these fits are included as red dots in the temperature–field (T – B) phase diagram of YbRh_2Si_2 (Fig. 3a). A linear fit to these points (dashed red line in Fig. 3a denoted T_{Hall}) extrapolates at zero temperature to the critical field $B_{1c} \approx 0.7$ T for the disappearance of antiferromagnetic order and the QCP. Thus, the crossover is directly related to the QCP. The sharpness of the crossover is best quantified by the full-width at half-maximum (FWHM) of $d\tilde{R}_H/dB_1$, which represents the change of slope of $\rho_H(B_1)$. The temperature dependence of the FWHM values is well described by a pure power law, $\text{FWHM} \propto T^a$, $a = 0.5 \pm 0.1$ (inset of Fig. 3a), suggesting that, at zero temperature, the crossover is infinitely sharp ($\text{FWHM} = 0$).

To understand the origin of this sharp feature, we now turn to the crossed-field measurement of the linear-response Hall coefficient, equation (1). The inset of Fig. 2b displays $\rho_H(B_1)$ curves taken at 65 mK for different values of the longitudinal tuning field B_2 . With increasing B_2 the linear-response Hall coefficient R_H decreases. For a quantitative analysis we fit, as above, $\int \tilde{R}_H(B)dB$ to the $\rho_H(B_1)$ data (solid lines in the inset of Fig. 2b). As opposed to the single-field experiment, $R_H^0 = R_H$ is now the only parameter to consider. R_H , normalized to its value at the crossover field B_0 , is plotted in the main panel of Fig. 2b as a function of the normalized tuning field B_2/B_0 . Data obtained in the same way at 45, 75 and 93 mK are included as well. R_H decreases as a function of B_2 by a factor of ~ 1.5 . In a simple one band model, this corresponds to an increase in the charge carrier concentration from ~ 2 to ~ 3 holes per YbRh_2Si_2 formula unit. The crossover sharpens up as the temperature is lowered. For a quantitative analysis, we may now fit the crossover form $R_H(B) = R_H^\infty - (R_H^\infty - R_H^0)\gamma(B)$ to the $R_H(B_2)$ data (solid curves in main panel of Fig. 2b). The R_H^∞ values obtained for these four temperatures are included as green triangles in the main part of Fig. 1a, showing that the Hall coefficient in the field-induced Landau Fermi liquid (LFL) state (compare Fig. 3a) at very low temperatures is substantially smaller than in the $B = 0$ antiferromagnetically ordered state. The $11B_0$ and FWHM values obtained from the above fits are included as green dots in Fig. 3a and its inset. The factor of 11 accounts for the fact that the tuning field B_2 is applied in the easy tetragonal plane of YbRh_2Si_2 where, owing to the magnetic anisotropy, the action of a magnetic field is known to be ~ 11 times as strong as along the magnetically hard c axis¹². For both quantities, the green and red data points agree within the error bars. Thus, the linear Hall response $R_H(B_2)$ of the crossed-field measurement and the differential Hall response $\tilde{R}_H(B_1)$ of the single-field measurement can be described by the same functional form, and the respective crossover positions and crossover widths agree quantitatively. This experimental finding suggests that the second term on the right hand side of equation (2) plays a minor role, at least in the experimentally accessed part of the T – B phase diagram. Therefore, here the single-field experiment appears to probe essentially the same (linear-response) Hall coefficient as the crossed-field experiment. However, there is a quantitative difference in the jump heights of $\tilde{R}_H(B_1)$ and $R_H(B_2)$, which probably reflects the anisotropies in the evolution of the electronic bandstructure under transverse and longitudinal field-tuning²¹, amplified by the likely presence of a multisheeted, anisotropic Fermi surface.

The phase diagram in the magnetic field–temperature parameter space can now be illustrated by a three-dimensional representation of $d\gamma(B)/dB$ (Fig. 3b). $\gamma(B)$ is calculated at arbitrary temperatures from the linear B_0 versus T fit (dashed red line in Fig. 3a) and a power law fit (not shown) to the $p(T)$ data obtained from the fits to $\rho_H(B_1)$ (Fig. 2a) and to $R_H(B_2)$ (main panel of Fig. 2b). With decreasing temperature, the $d\gamma(B)/dB$ curves sharpen up and their crossover position B_0 , designated by drop lines, shifts to lower fields such that, at zero temperature, a δ -function (dashed line in the

$T = 0$ plane in Fig. 3b) is situated at the QCP.

Thus, the extrapolation of our finite temperature data to zero temperature indicates the presence of a finite discontinuity ('jump') in the Hall coefficient at the QCP, even though the change in the magnetic order parameter is infinitesimal¹¹. By contrast, in an itinerant SDW scenario, the Fermi surface is expected⁷ to fold over at the QCP; the Hall coefficient is then continuous across the QCP, evolving gradually with the size of the antiferromagnetic order parameter, as is indeed observed experimentally²². (For a more quantitative comparison, see Fig. 2b legend). Our results hint at a sudden reconstruction of the Fermi surface at the QCP, corresponding to the sudden loss of 'mobile' $4f$ electrons^{7,8,16}. Loosely speaking, the volume of the Fermi surface has changed discontinuously. Here of course, the concept of Fermi surface volume needs to be treated with some care, for antiferromagnetism breaks translational symmetry. YbRh_2Si_2 may well be an easy-plane incommensurate antiferromagnet, and for this class of antiferromagnet, to linear order in the magnetic order parameter, the Fermi surface volume is well-defined in the paramagnetic unit cell. From our data we infer that the antiferromagnetic ground state has a 'small' Fermi surface, which is the same as the one extracted from the high-temperature Hall effect data (main panel of Fig. 1a), while the paramagnetic ground state has a 'large' Fermi surface, which presumably incorporates the new heavy-fermion states injected by the local moments.

The crossover line $T_{\text{Hall}}(B)$ (Fig. 3) is then interpreted as the finite temperature signature of the 'jump' in the Fermi surface volume. It delineates the position at which a new large Fermi surface emerges in the incoherent electron fluid. It is interesting that this precursor to heavy quasiparticle formation takes place at temperatures well above the temperature T^* (dashed black curve in Fig. 3a) below which the coherent LFL develops. The existence of a large Fermi surface in the absence of well-defined quasiparticles is well known in one-dimensional Luttinger liquids²³. It is also reminiscent of the marginal Fermi liquid behaviour in copper oxide superconductors, where a large Fermi surface is seen in the angle-resolved photoemission studies, but the scattering rate grows linearly, rather than quadratically, in temperature^{24,25}. Note also that the crossover line $T_{\text{Hall}}(B)$ does not follow the antiferromagnetic transition line $T_N(B)$ (Fig. 3a). Indeed, within experimental resolution, the initial Hall coefficient shows no change at the zero field Néel temperature of 70 mK (Fig. 1a). This behaviour contrasts markedly with that expected in an itinerant SDW, where changes in the Hall coefficient should coincide with the Néel transition—as is indeed observed for $\text{Cr}_{1-x}\text{V}_x$ (refs 22 and 26). Thus we may discard the possibility that the observed crossover in the Hall coefficient of YbRh_2Si_2 is due to a unit-cell doubling in a symmetry breaking antiferromagnetic transition.

Even though the crossover at $T_{\text{Hall}}(B)$ broadens rapidly with temperature (compare $\text{FWHM}(T)$ in the inset of Fig. 3a and width of $d\gamma(B)/dB$ in Fig. 3b), so that it can not be followed beyond about 0.5 K, the additional contribution ΔR_H to the initial Hall coefficient (main panel of Fig. 1a), which we attribute to fluctuations of the Fermi surface volume, can be discerned up to much higher temperatures of the order of 10 K. This is precisely the temperature below which NFL behaviour is observed in thermodynamic and dynamical properties^{9,12}. This observation makes it very tempting to hold fluctuations of the Fermi surface volume responsible for the NFL behaviour observed over this same temperature window. The fact that the NFL behaviour is observed in the entire phase diagram above T_N and T^* (and below 10 K) can be related to the broadness of the crossover. Interestingly, the spin fluctuation scale T_0 (extracted from a logarithmic fit $\Delta C_p/T \propto \ln(T_0/T)$ to the specific heat data for $0.3 \text{ K} < T < 10 \text{ K}$) and the single-ion Kondo temperature (which marks the onset of magnetic screening) extracted from a magnetic entropy measurement are of the same order of magnitude^{9,12}.

To summarize, we observe a rapid crossover of the Hall coefficient

as a function of a control parameter. By extrapolation to $T = 0$ of both the Hall crossover and the magnetic phase transition¹², we infer that a large jump of the Hall coefficient occurs at the QCP. We expect this new insight, made possible primarily by the absence of superconductivity, to have broad implications for other strongly correlated electron systems²⁷. □

Methods

Anomalous Hall effect

In general, the Hall effect of materials containing localized magnetic moments is dominated at high temperatures by an anomalous Hall effect produced by the left-right asymmetry in incoherent electron scattering processes²⁸. The initial or linear-response Hall coefficient R_H (Hall coefficient in zero-field limit) scales for many materials with the product of electrical resistivity ρ and magnetic susceptibility χ ,

$$R_H = R_0 + C\rho\chi \quad (3)$$

where R_0 is the normal Hall coefficient and C is a constant²⁸. The term $C\rho\chi$ represents the anomalous Hall effect due to intrinsic scattering. The temperature-independent extrinsic anomalous Hall coefficient R_{ex} due to skew scattering by residual defects may be estimated from

$$R_{ex} = C\rho_0\chi_0 \quad (4)$$

where ρ_0 is the residual resistivity and χ_0 the residual volume magnetic susceptibility²⁸. A model including crystalline electric field effects valid in the incoherent regime²⁹, on the other hand, predicts

$$R_H = R_0 + R_s\chi \quad (5)$$

instead of equation (3). Here R_s is a constant and $R_s\chi$ the anomalous Hall-effect term.

In Fig. 1a we have shown that also in YbRh₂Si₂ the high-temperature Hall coefficient is dominated by the anomalous Hall effect. Between 7 and 300 K (90 and 300 K), equation (3) (equation (5)) holds (compare inset of Fig. 1a). The R_0 value obtained for both models is $(2.4 \pm 0.1) \times 10^{-10} \text{ m}^3 \text{ C}^{-1}$, which corresponds, in a simple one band model, to a charge carrier concentration of $2.6 \times 10^{28} \text{ m}^{-3}$ (approximately 2 holes per formula unit of YbRh₂Si₂). Considering only the magnetic contribution to ρ in equation (3) yields similar values for R_0 (ref. 18). Below about 1 K, where the extrapolation of the fit according to equation (3) (red dashed curve in Fig. 1a) becomes temperature independent and saturates at the value of the normal Hall coefficient R_0 , the intrinsic anomalous Hall effect is negligible. The extrinsic anomalous Hall effect estimated from equation (4) with $\rho_0 \approx 1 \mu\Omega \text{ cm}$ and $\chi_0 = 0.0035 \text{ (B/C, } T = 40 \text{ mK)}$ (ref. 12) is less than 4% of R_0 and thus plays a negligible role. Therefore, below about 1 K, the initial Hall coefficient of YbRh₂Si₂ is essentially free of any anomalous contribution.

The anomalous Hall effect in finite magnetic fields may, in analogy with equation (3), be estimated from

$$\rho_{H,a}(B) = C\rho(B)\mu_0 M(B) \quad (6)$$

where $\rho(B)$ and $M(B)$ are the field-dependent electrical resistivity and magnetization, respectively.

For YbRh₂Si₂, $\rho(B)$ (not shown) and $M(B)$ (ref. 12) have been measured in the relevant geometry ($B \parallel c$, current $I \perp c$). For the parameter C we use the value extracted from the temperature dependence of the initial Hall coefficient (inset of Fig. 1a). $\rho_{H,a}$ is less than 20% of ρ_H at all temperatures and fields.

Received 1 July; accepted 14 October 2004; doi:10.1038/nature03129.

- Martin, R. M. Fermi-surface sum rule and its consequences for periodic Kondo and mixed-valence systems. *Phys. Rev. Lett.* **48**, 362–365 (1982).
- Fulde, P. in *Narrow-Band Phenomena—Influence of Electrons with both Band and Localized Character* (ed. Fuggle, J. C.) 27–29 (Plenum, New York, 1988).
- Oshikawa, M. Topological approach to Luttinger's theorem and Fermi surface of a Kondo lattice. *Phys. Rev. Lett.* **84**, 3370–3373 (2000).
- Hertz, J. A. Quantum critical phenomena. *Phys. Rev. B* **14**, 1165–1184 (1976).
- Millis, A. J. Effect of a nonzero temperature on quantum critical points in itinerant fermion systems. *Phys. Rev. B* **48**, 7183–7196 (1993).
- Schröder, A. et al. Onset of antiferromagnetism in heavy-fermion metals. *Nature* **407**, 351–355 (2000).
- Coleman, P., Pépin, C., Si, Q. & Ramazashvili, R. How do Fermi liquids get heavy and die? *J. Phys. Condens. Matter* **13**, R723–R738 (2001).
- Si, Q., Rabello, S., Ingersent, K. & Smith, J. L. Locally critical quantum phase transitions in strongly correlated metals. *Nature* **413**, 804–808 (2001).
- Trovarelli, O. et al. YbRh₂Si₂: Pronounced non-Fermi-liquid effects above a low-lying magnetic phase transition. *Phys. Rev. Lett.* **85**, 626–629 (2000).
- Ishida, K. et al. YbRh₂Si₂: Spin fluctuations in the vicinity of a quantum critical point at low magnetic fields. *Phys. Rev. Lett.* **89**, 107202 (2002).
- Ishida, K. et al. Low-temperature magnetic order and spin dynamics in YbRh₂Si₂. *Phys. Rev. B* **68**, 184401 (2003).
- Gegenwart, P. et al. Magnetic-field induced quantum critical point in YbRh₂Si₂. *Phys. Rev. Lett.* **89**, 056402 (2002).
- Küchler, R. et al. Low-temperature thermal expansion and magnetostriction of YbRh₂(Si_{1-x}Ge_x)₂ ($x = 0$ and 0.05). *J. Magn. Magn. Mater.* **272–276**, 229–230 (2004).
- Custers, J. et al. The break-up of heavy electrons at a quantum critical point. *Nature* **424**, 524–527 (2003).
- Stewart, G. R. Non-Fermi-liquid behavior in d - and f -electron metals. *Rev. Mod. Phys.* **73**, 797–855 (2001).

- Senthil, T., Vojta, M. & Sachdev, S. Weak magnetism and non-Fermi liquids near heavy-fermion critical points. *Phys. Rev. B* **69**, 035111 (2004).
- Paschen, S. et al. Hall effect of the NFL compound YbRh₂Si₂. *Acta Phys. Polon.* **B 34**, 359–362 (2003).
- Paschen, S. et al. Anomalous Hall effect in YbRh₂Si₂. *Physica B* (in the press).
- Anderson, P. W. Hall effect in the two-dimensional Luttinger liquid. *Phys. Rev. Lett.* **67**, 2092–2094 (1991).
- Nakajima, Y. et al. Normal-state Hall angle and magnetoresistance in quasi-2D heavy fermion CeCoIn₅ near a quantum critical point. *J. Phys. Soc. Jpn* **73**, 5–8 (2004).
- Custers, J. *Quantum-Critical Behavior in the Heavy-Fermion Compounds YbRh₂Si₂ and CeIn_{3-x}Sn_x*. PhD thesis, TU Dresden (2004).
- Lee, M., Husmann, A., Rosenbaum, T. F. & Aeppli, G. High resolution study of magnetic ordering at absolute zero. *Phys. Rev. Lett.* **92**, 187201 (2004).
- Haldane, F. D. M. 'Luttinger liquid theory' of one-dimensional quantum fluids: I. Properties of the Luttinger model and their extension to the general 1D interacting spinless Fermi gas. *J. Phys. C* **14**, 2585–2609 (1981).
- Varma, C. M., Littlewood, P. B., Schmitt-Rink, S., Abrahams, E. & Ruckenstein, A. E. Phenomenology of the normal state of Cu-O high-temperature superconductors. *Phys. Rev. Lett.* **63**, 1996–1999 (1989).
- Valla, T. et al. Evidence for quantum critical behaviour in the optimally doped cuprate Bi₂Sr₂CaCu₂O_{8+δ}. *Science* **285**, 2110–2113 (1999).
- Yeh, A. et al. Quantum phase transition in a common metal. *Nature* **419**, 459–462 (2002).
- Balakirev, F. F. et al. Signature of optimal doping in Hall-effect measurements on a high-temperature superconductor. *Nature* **424**, 912–915 (2003).
- Fert, A. & Levy, P. M. Theory of the Hall effect in heavy-fermion compounds. *Phys. Rev. B* **36**, 1907–1916 (1987).
- Kontani, H., Miyazawa, M. & Yamada, K. Theory of anomalous Hall effect in a heavy fermion system with a strong anisotropic crystal field. *J. Phys. Soc. Jpn* **66**, 2252–2255 (1997).
- Norman, M. R., Si, Q., Bazaliy, Y. B. & Ramazashvili, R. Hall effect in nested antiferromagnets near the quantum critical point. *Phys. Rev. Lett.* **90**, 116601 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We acknowledge discussions with P. Bühler, J. Custers, C. Langhammer, C. Pépin, A. Rosch, A. Schofield, M. Vojta, A. Tselik and F. Weickert. Part of the work at Dresden was supported by the Fonds der Chemischen Industrie. P.C. and Q.S. are supported by the National Science Foundation. The work at Rice University was partially supported by the Welch Foundation and TCSAM.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.P. (paschen@cpfs.mpg.de).

Lensless imaging of magnetic nanostructures by X-ray spectro-holography

S. Eisebitt¹, J. Lüning², W. F. Schlotter^{2,3}, M. Lörger¹, O. Hellwig^{1,4}, W. Eberhardt¹ & J. Stöhr²

¹BESSY mbH, Albert-Einstein-Straße 15, 12489 Berlin, Germany

²SSL, Stanford Linear Accelerator Center, 2575 Sand Hill Road, Menlo Park, California 94025, USA

³Department of Applied Physics, 316 Via Pueblo Mall, Stanford University, Stanford, California 94305-4090, USA

⁴San Jose Research Center, Hitachi Global Storage Technologies, 650 Harry Road, San Jose, California 95120, USA

Our knowledge of the structure of matter is largely based on X-ray diffraction studies of periodic structures and the successful transformation (inversion) of the diffraction patterns into real-space atomic maps. But the determination of non-periodic nanoscale structures by X-rays is much more difficult. Inversion of the measured diffuse X-ray intensity patterns suffers from the intrinsic loss of phase information^{1,2}, and direct imaging methods are limited in resolution by the available X-ray optics³. Here we demonstrate a versatile technique for imaging nanostructures, based on the use of resonantly tuned soft X-rays for scattering contrast and the direct Fourier inversion of a holo-

graphically formed interference pattern. Our implementation places the sample behind a lithographically manufactured mask with a micrometre-sized sample aperture and a nanometre-sized hole that defines a reference beam. As an example, we have used the resonant X-ray magnetic circular dichroism effect to image the random magnetic domain structure in a Co/Pt multilayer film with a spatial resolution of 50 nm. Our technique, which is a form of Fourier transform holography, is transferable to a wide variety of specimens, appears scalable to diffraction-limited resolution, and is well suited for ultrafast single-shot imaging with coherent X-ray free-electron laser sources⁴.

Our indirect imaging technique is based on the use of tunable soft X-rays and simple Fourier inversion of a reciprocal space interference pattern to yield a real-space image. By tuning the energy of the X-rays to characteristic absorption edges we can generate contrast for different properties of the nanostructured sample, such as elemental and chemical composition. With further use of variable X-ray polarization the sensitivity and contrast is extended to charge⁵ and spin orientations^{6,7}. Our holography-based method overcomes the central problem of all indirect methods—the recovery of the phase information^{2,8,9}—with the use of a nanoscale reference aperture next to the sample that phases the recorded interference pattern^{1,10}. Our method is simple and suitable for imaging a variety of nanostructures, and the achievable resolution is, in principle, limited only by the soft X-ray wavelength, which is about a factor of ten smaller than the resolution achievable with image-forming X-ray lenses today.

This lensless spectro-holography technique does not require any kind of image-forming lenses, such as zone plates used in conventional X-ray microscopy¹¹ or electron optical lenses used for image formation in X-ray-induced photoemission electron microscopy¹². Because holography-based methods require a coherence length that is longer than the optical path length differences encountered in the experiment, they have largely been restricted to lasers. For nanoscale applications, however, shorter-wavelength radiation is needed to overcome the diffraction-limited spatial resolution of conventional lasers¹³. In principle, X-rays allow higher spatial resolution, but in practice, submicrometre resolution has rarely been achieved¹⁴, with

the notable exceptions of refs 15 and 16, in which spatial resolutions of about 50 nm were achieved using combined holography and microscopy approaches^{15,16}.

Our studies use the high coherent flux and the polarization and photon-energy tunability available at advanced synchrotron radiation sources, such as the BESSY-II storage ring, where we carried out the present work. In particular, we use the large X-ray magnetic circular dichroism effect at the L-edges of the transition metals for image contrast¹⁷ so that our method is a true combination of spectroscopy and holography. The new technique is complementary to direct magnetic X-ray imaging methods based on photoemission electron microscopy or transmission X-ray microscopy.

Our experimental arrangement is illustrated in Fig. 1. We used circularly polarized soft X-rays from an undulator source in conjunction with a spherical grating monochromator that determined the longitudinal coherence length. At the photon energy 778 eV (wavelength 1.59 nm) used in our study, corresponding to the Co L₃ absorption edge, the longitudinal coherence length was $\xi_l = \lambda^2 / (2\Delta\lambda) = 1.6 \mu\text{m}$, where λ is the wavelength. The beam from the monochromator was incident on an aperture of diameter $D = 20 \mu\text{m}$ that acted as a transverse coherence filter. The central Airy disk of the transmitted beam coherently illuminated a nanostructured transmission mask directly in front of the sample, both of which are placed $z = 723 \text{ mm}$ behind the coherence filter. This geometry determined the transverse coherence length of $\xi_t = (\lambda z) / (2\pi D) = 9.1 \mu\text{m}$. An in-vacuum charge-coupled device (CCD) camera, positioned 315 mm downstream of the mask-sample structure, was used to record the hologram.

The key element in this experiment is the mask-sample arrangement. We chose an integrated mask-sample design, as illustrated in the lower inset of Fig. 1. The sample-mask structure was fabricated by use of a Si₃N₄ membrane on a Si support frame. On one side of the membrane we sputtered a 600-nm-thick gold film. Unlike the Si₃N₄ membrane, the thicker gold film is opaque for 778-eV X-rays. On the other side of the membrane a magnetic multilayer with 50 repeats of Co(4 Å)/Pt(7 Å) bilayers was sputter-deposited on a 20-nm-thick Pt base layer¹⁸. The multilayer was capped with a 2-nm Pt layer to prevent corrosion. We then used a focused ion beam to cut a circular aperture of diameter 1.5 μm out of the gold film, down to the Si₃N₄ membrane.

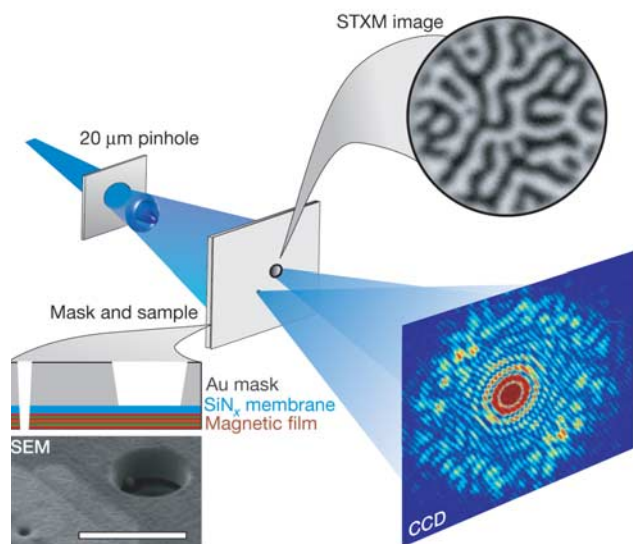


Figure 1 Scheme of the experimental set-up. Monochromatized and circular polarized X-rays are incident on a mask-sample structure after spatial coherence filtering. The object and reference beam are defined by the mask, and the resulting hologram is recorded on a CCD detector. The lower inset shows the geometry and an electron microscopy image of the mask-sample structure. The scale bar in the microscopy image is 2.0 μm . The top inset shows a STXM image of the magnetic structure illuminated through the sample aperture. The field of view is 1.5 μm .

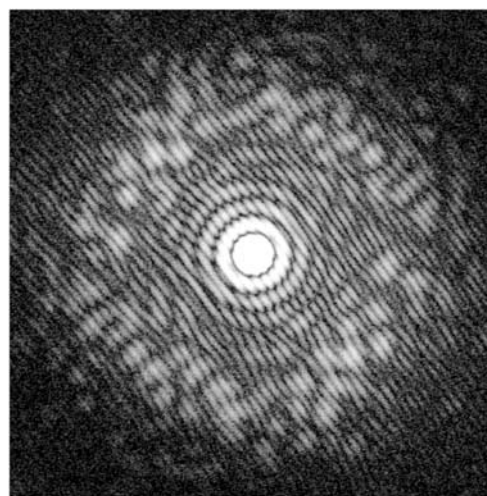


Figure 2 Hologram recorded with X-rays (right circular polarization) at a wavelength of 1.59 nm. The maximum in-plane momentum transfer in the measurement is $\pm 0.13 \text{ nm}^{-1}$, shown up to $\pm 0.06 \text{ nm}^{-1}$ in the image. Intensity is represented on a logarithmic grey scale, with black denoting the minimum intensity of 10^3 and white denoting the maximum intensity of 10^5 . Black and white appear saturated in the picture only, the dynamic range of the hologram is 10^7 .

This 'sample aperture' defines the object beam through the sample. An image of the magnetic domain structure within the sample aperture was recorded by scanning transmission X-ray microscopy¹⁹ (STXM) and is shown in the upper inset of Fig. 1. It reveals a characteristic worm domain pattern where the black and white areas correspond to domains that have opposite out-of-plane magnetization directions¹⁸. Next to the sample aperture, at a centre-to-centre distance of 3 μm , a circular pinhole was drilled by a focused ion beam through the entire mask-sample structure. This high-aspect-ratio pinhole defines the reference beam. It had a conical shape along the beam direction with a diameter of 350 nm at the (Au) front side and 100 nm at the (multilayer) back side.

The in-plane reference hole and sample aperture conveniently define a lensless Fourier transform holography geometry^{20,21}. The Fourier transform hologram recorded with circularly polarized soft

X-rays of positive helicity is shown in Fig. 2. The incident intensity was adjusted so that the intense central peak in the hologram could be detected without saturation down to a minimum momentum transfer of $2\pi/(10\text{ }\mu\text{m})$. This led to typical exposure times of 10 s per frame. Fifty frames were accumulated to improve the counting statistics in the image. The Fraunhofer ring pattern dominating the centre of the hologram originates from the sample aperture with diameter 1.5 μm . A high-frequency oscillation created by the interference of object and reference wave is visible with high contrast. Small angle scattering caused by the magnetic domain structure^{17,22,23} in the Co/Pt multilayer film is visible as a broad ring with an average momentum transfer of $|q| = 0.036\text{ nm}^{-1}$. This small angle scattering ring is broken up into individual speckle attributable to object-object and object-reference interference.

Owing to the off-axis geometry in Fourier transform holography, the object image and its conjugate can easily be separated and retrieved by a single Fourier transformation of the scattering intensity¹. In Fig. 3a we present a digital two-dimensional fast Fourier transformation of the magnetic X-ray hologram in Fig. 2. The intense area in the centre contains the sample-sample and reference hole-reference hole autocorrelations. The sample-reference hole cross-correlations are seen on opposite sides of the central structure. They contain the desired information and directly give a high-quality transmission image of the lateral magnetic domain structure. On helicity reversal of the incident soft X-rays, we observe in the image an inversion of the magnetic domain contrast as expected, while the mask contrast remains unaffected²³. Therefore, as in magnetic spectro-microscopy, the difference of two opposite helicity images can be used to enhance the magnetic contrast and suppress any non-magnetic contributions. A helicity difference image is shown in Fig. 3b.

The magnetization map obtained by X-ray holography is in perfect agreement with the one obtained by STXM (Fig. 1, top right). Note that both have been recorded by exploiting the same dichroism contrast mechanism. This provides unambiguous proof that our lensless X-ray Fourier transform holography approach gives a unique real-space image of the sample.

To evaluate the spatial resolution, line scans through the magnetic profile obtained with spectro-microscopy and spectro-holography are compared in Fig. 3b. Both profiles are in quantitative agreement, indicating that the lateral resolution is essentially the same. The 10% to 90% contrast change in the image profiles has a width of 50 nm. The high spatial resolution is somewhat surprising, because in Fourier transform holography it is limited by the size of the reference source given in our case by the 100-nm diameter of the reference pinhole. Scanning electron microscopy (SEM) images show only the entrance and exit of the pinhole, so we cannot rule out the presence of a smaller constriction inside the high aspect ratio (1:8) pinhole channel. It is more likely, however, that the pinhole, tapered because of the 8° convergence of the focused ion beam, acts as a capillary waveguide and reduces the focal spot size²⁴.

Our experimental scheme, based on a nanostructured transmission mask on a transparent membrane, can be applied to a variety of different specimens. Samples can be grown directly on the back of a membrane, as presented here, or be introduced on a separate membrane and mounted in direct contact with the mask. Nanoparticles, colloids or biological samples could be placed into the object area in a controlled fashion by using, for example, integrated light microscopy and manipulator stages, as used for microinjection in biotechnology. Furthermore, one can easily envision an entire array of sample and reference apertures, allowing high-throughput investigations.

Our fixed arrangement of sample and reference aperture within the same plane greatly facilitates alignment, reduces the sensitivity to beam drift and eliminates focusing corrections in the holographic image reconstruction. It is also possible to record images with different well-defined shapes of the reference hole to apply appro-

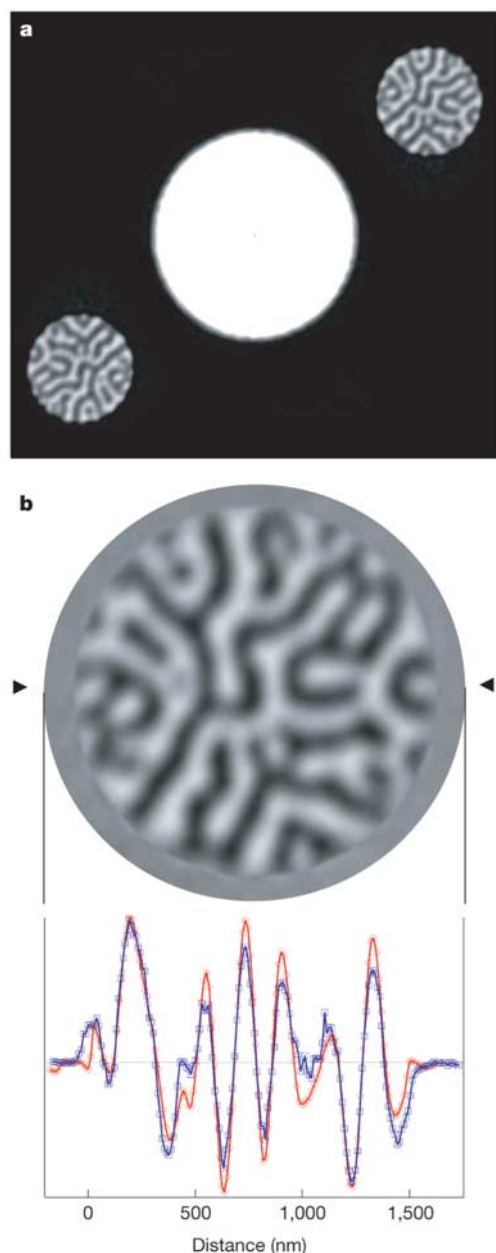


Figure 3 Images retrieved from the hologram. **a**, Two-dimensional fast Fourier transformation of the hologram in Fig. 2. **b**, Zoomed-in image, obtained by subtracting the Fourier transformations of opposite-helicity holograms. Below are shown scan lines through the holographic image (red) and the STXM image in Fig. 1 (blue).

priate source-shape corrections. The spatial resolution can be improved by using smaller apertures and through waveguide focusing effects, if it is not intensity-limited by the detectable maximum momentum transfer. When reducing the size of the reference aperture, either the X-ray transmission of the sample or the sample aperture size may have to be reduced to control the relative sample and reference beam intensities. If a reference beam intensity larger than the sample beam intensity cannot be achieved, nonlinearities in the holographic imaging process will have to be taken into account. A relative drift of the mask-sample structure relative to the CCD detector is not critical as long as it is smaller than the CCD pixel size (13.5 μm). Vibration isolation and thermal stabilization is therefore easy in comparison with other submicrometre microscopy techniques. Furthermore, extreme sample environments such as high magnetic fields or low temperatures can easily be implemented because there are no stringent space constraints around the sample.

Another exciting possibility is an improvement in spatial resolution by application of iterative phase retrieval methods^{2,8,9,25} to the hologram. Our mask-based X-ray Fourier transform holography approach is fully compatible with iterative phase retrieval, provided that the hologram is recorded with sufficient resolution. In a combined approach we record a high-resolution X-ray hologram and the holographic image obtained by Fourier transformation is then used as the starting point for iterative refinement of the phase. In this way, ambiguities in the iterative process are restricted to the finest details of the image below the resolution of the holographic image. The ultimate resolution is then determined not by the reference pinhole size but by the maximum photon momentum transfer or, eventually, by the wavelength.

The present work was in part motivated by the expected availability of X-ray free-electron lasers in the near future^{26,27}. Such sources will provide coherent high-intensity X-ray pulses of femtosecond duration. From the exposure time and coherent flux used for our images we can estimate that a single X-ray free-electron laser pulse will be sufficient to record an ultrafast single shot image. This will open the door for taking ultrafast movies of processes on the nanometre length scale, for example, of phase transitions. Our holographic approach is well matched to the spatial and temporal source properties of a free-electron laser and is the method of choice for single-shot ultrafast imaging with such sources⁴. □

Methods

Soft X-rays with circular polarization are generated in a APPLE-II-type helical undulator with 30 periods of length 56 mm. An energy resolution of $\lambda/\Delta\lambda = 2,000$ was obtained with a spherical grating monochromator. The 20- μm diameter aperture used for spatial filtering was located 6 m downstream of the beamline focus, thus accepting approximately 0.1% of the total beamline photon flux. Holograms were recorded with a Princeton Instruments PI-SX CCD camera equipped with an in-vacuum back-illuminated chip of $2,048 \times 2,048$ pixels of width 13.5 μm . For the hologram shown in Fig. 2, about 1×10^{10} photons were incident on the CCD chip during the accumulated exposure time. The apertures of the Fourier transform holography mask were prepared with a dual-beam focused ion beam instrument from FEI (Strata 235). The scanning X-ray microscopy image was recorded on the microscopy station on beamline 11 at the Advanced Light Source, Berkeley. The microscope has been described by ref. 19.

Received 8 July; accepted 28 October 2004; doi:10.1038/nature03139.

1. Stroke, G. W. *An Introduction to Coherent Optics and Holography* (Academic, New York, 1969).
2. Miao, J. W., Charalambous, P., Kirz, J. & Sayre, D. Extending the methodology of X-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens. *Nature* **400**, 342–344 (1999).
3. Chao, W. L. *et al.* 20-nm-resolution soft x-ray microscopy demonstrated by use of multilayer test structures. *Opt. Lett.* **28**, 2019–2021 (2003).
4. Neutze, R., Wouts, R., Van der Spoel, D., Weckert, E. & Hajdu, J. Potential for biomolecular imaging with femtosecond X-ray pulses. *Nature* **406**, 752–757 (2000).
5. Ade, H. & Hsiao, B. X-ray linear dichroism microscopy. *Science* **262**, 1427–1429 (1993).
6. Stöhr, J. *et al.* Element-specific magnetic microscopy with circularly polarized X-rays. *Science* **259**, 658–661 (1993).
7. Fischer, P. *et al.* Magnetic domain imaging with a transmission X-ray microscope. *J. Magn. Magn. Mater.* **199**, 624–627 (1999).
8. Gerchberg, R. W. & Saxton, W. O. Practical algorithm for determination of phase from image and diffraction plane pictures. *Optik* **35**, 237–246 (1972).
9. Fienup, J. R. Phase retrieval algorithms—a comparison. *Appl. Opt.* **21**, 2758–2769 (1982).

10. Gabor, D., Kock, W. E. & Stroke, G. W. Holography. *Science* **173**, 11–23 (1971).
11. Focus on X-ray microscopy. *Synchrotron Radiat. News* (special issue, ed. Schmahl, G.) **16**(3), 2–63 (2003).
12. Stöhr, J. *et al.* Element-specific magnetic microscopy with circularly polarized X-rays. *Science* **259**, 658–661 (1993).
13. Xu, W. B., Jericho, M. H., Meinertzhagen, I. A. & Kreuzer, H. J. Digital in-line holography for biological applications. *Proc. Natl Acad. Sci. USA* **98**, 11301–11305 (2001).
14. Trebes, J. E. *et al.* Demonstration of X-ray holography with an X-ray laser. *Science* **238**, 517–519 (1987).
15. McNulty, I. *et al.* High-resolution imaging by Fourier-transform X-ray holography. *Science* **256**, 1009–1012 (1992).
16. Lindaas, S., Howells, H., Jacobsen, C. & Kalinovsky, A. X-ray holographic microscopy by means of photoresist recording and atomic-force microscope readout. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* **13**, 1788–1800 (1996).
17. Hannon, J. P., Trammell, G. T., Blume, M. & Gibbs, D. X-ray resonance exchange scattering. *Phys. Rev. Lett.* **61**, 1245–1248 (1988).
18. Hellwig, O., Denbeaux, G. P., Kortright, J. B. & Fullerton, E. E. X-ray studies of aligned magnetic stripe domains in perpendicular multilayers. *Physica B Condens. Matter* **336**, 136–144 (2003).
19. Kilcoyne, A. L. D. *et al.* Interferometer-controlled scanning transmission X-ray microscopes at the Advanced Light Source. *J. Synchrotron Radiat.* **10**, 125–136 (2003).
20. Stroke, G. W. Lensless Fourier-transform method for optical holography. *Appl. Phys. Lett.* **6**, 201–203 (1965).
21. Winthrop, J. T. & Worthing, C. R. X-ray microscopy by successive Fourier transformation. *Phys. Lett.* **15**, 124–126 (1965).
22. Kortright, J. B. *et al.* Soft-x-ray small-angle scattering as a sensitive probe of magnetic and charge heterogeneity. *Phys. Rev. B* **64**, 92401 (2001).
23. Eisebitt, S. *et al.* Polarization effects in coherent scattering from magnetic specimen: Implications for x-ray holography, lensless imaging, and correlation spectroscopy. *Phys. Rev. B* **68**, 104419 (2003).
24. Bergemann, C., Keymeulen, H. & Van der Veen, J. F. Focusing X-ray beams to nanometer dimensions. *Phys. Rev. Lett.* **91**, 204801 (2003).
25. Fienup, J. R. Reconstruction of a complex-valued object from the modulus of its Fourier transform using a support constraint. *J. Opt. Soc. Am.* **4**, 118–123 (1987).
26. Cho, A. The ultimate bright idea. *Science* **296**, 1008–1010 (2002).
27. Emma, P. *et al.* Femtosecond and subfemtosecond X-ray pulses from a self-amplified spontaneous emission-based free-electron laser. *Phys. Rev. Lett.* **92**, 074801 (2004).

Acknowledgements We thank Y. Acremann for recording the STXM image of the sample at the Advanced Light Source and E. E. Fullerton for access to his thin film deposition facility. The work of the SSRL authors is supported by the US Department of Energy, Office of Basic Energy Sciences.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.E. (eisebitt@bessy.de) or J.L. (luning@stanford.edu).

High temperatures in the Late Cretaceous Arctic Ocean

Hugh C. Jenkyns¹, Astrid Forster², Stefan Schouten² & Jaap S. Sinninghe Damsté²

¹Department of Earth Sciences, University of Oxford, Parks Road, Oxford OX1 3PR, UK

²Royal Netherlands Institute for Sea Research (NIOZ), Department of Marine Biogeochemistry and Toxicology, PO Box 59, 1790 AB Den Burg, Texel, The Netherlands

To understand the climate dynamics of the warm, equable greenhouse world of the Late Cretaceous period, it is important to determine polar palaeotemperatures. The early palaeoceanographic history of the Arctic Ocean has, however, remained largely unknown, because the sea floor and underlying deposits are usually inaccessible beneath a cover of floating ice. A shallow piston core taken from a drifting ice island in 1970 fortuitously retrieved unconsolidated Upper Cretaceous organic-rich sediment from Alpha ridge^{1–4}, a submarine elevated feature of probable oceanic origin⁵. A lack of carbonate in the sediments from this core has prevented the use of traditional oxygen-isotope palaeothermometry. Here we determine Arctic palaeotemperatures from these Upper Cretaceous deposits using TEX₈₆,

a new palaeothermometer that is based on the composition of membrane lipids derived from a ubiquitous component of marine plankton, *Crenarchaeota*⁶. From these analyses we infer an average sea surface temperature of ~15 °C for the Arctic Ocean about 70 million years ago⁷. This calibration point implies an Equator-to-pole gradient in sea surface temperatures of ~15 °C during this interval and, by extrapolation, we suggest that polar waters were generally warmer than 20 °C during the middle Cretaceous (~90 million years ago).

Only three cores containing well-dated Cretaceous deep-sea sediment are known from the Arctic Ocean. The piston cores FL-437 and FL-533 were taken during the drift of the Arctic ice island T-3 (Fletcher's ice island) over Alpha ridge (Fig. 1) during the period 1963–74. FL-437, cored in 1969, comprises yellowish laminated siliceous ooze rich in exquisitely preserved diatoms (*Anaulus*, *Hemiaulus* and *Rhizosolenia* being the most abundant genera in a low-diversity assemblage), with subsidiary ebridians, silicoflagellates, archaeomonads and sparse fish remains: the initial age attribution of mid- to late Maastrichtian¹ has now been extended to include the Campanian⁸. Core 6 of the Canadian Expedition to Study the Alpha Ridge (CESAR), taken in 1983 from a research station installed on drifting sea ice (Fig. 1), is composed of siliceous sediments very similar to those of core FL-437, being spectacularly laminated on a millimetre scale and similarly lacking planktonic carbonate^{1,3,9}. Total organic carbon (TOC) contents in these sediments are low, typically <1% (ref. 3). The age attribution of this core includes the interval Campanian–Maastrichtian, depending on

whether the diatoms, the silicoflagellates or the palynomorphs are taken as the prime biostratigraphic indicator^{10–12}.

FL-533, cored in 1970, comprises homogeneous (apart from a finely laminated uppermost 2 cm) organic-rich black mud whose TOC content ranges between 11.0 and 15.8 wt% (Fig. 2). The organic matter comprises amorphous material, abundant woody fragments, leaf cuticles, spores and pollen; marine dinoflagellates are particularly abundant². Hydrogen indices (150–350 mg hydrocarbons per g organic carbon) equally suggest the presence of marine and terrestrial components, a conclusion consistent with the presence of marine and terrestrial biomarkers. Our analyses indicate that biomarkers derived from diatoms are abundant in these deposits even though their siliceous frustules are not recorded, presumably because they have dissolved within the sediment: the C₂₅ highly branched isoprenoid is derived from centric diatoms of the order *Rhizosolenia* and indicates an age younger than Turonian, the period in which these forms evolved¹³. The relative abundance of 24-norcholestanes, a group of steroids, is also diagnostic because their occurrence is generally associated with high-latitude deposition of diatomaceous ooze¹⁴. Initially dated as late Campanian or late Campanian to Maastrichtian^{1,2}, re-examination of the dinoflagellate cysts, together with acritarchs and prasinophytes, suggests an early Maastrichtian age for these black muds⁴.

Integration of these various biostratigraphic indices from all three Upper Cretaceous sections allows more than one interpretation: the organic-rich and siliceous sediments could either be coeval or represent successive episodes of pelagic deposition (Fig. 2).

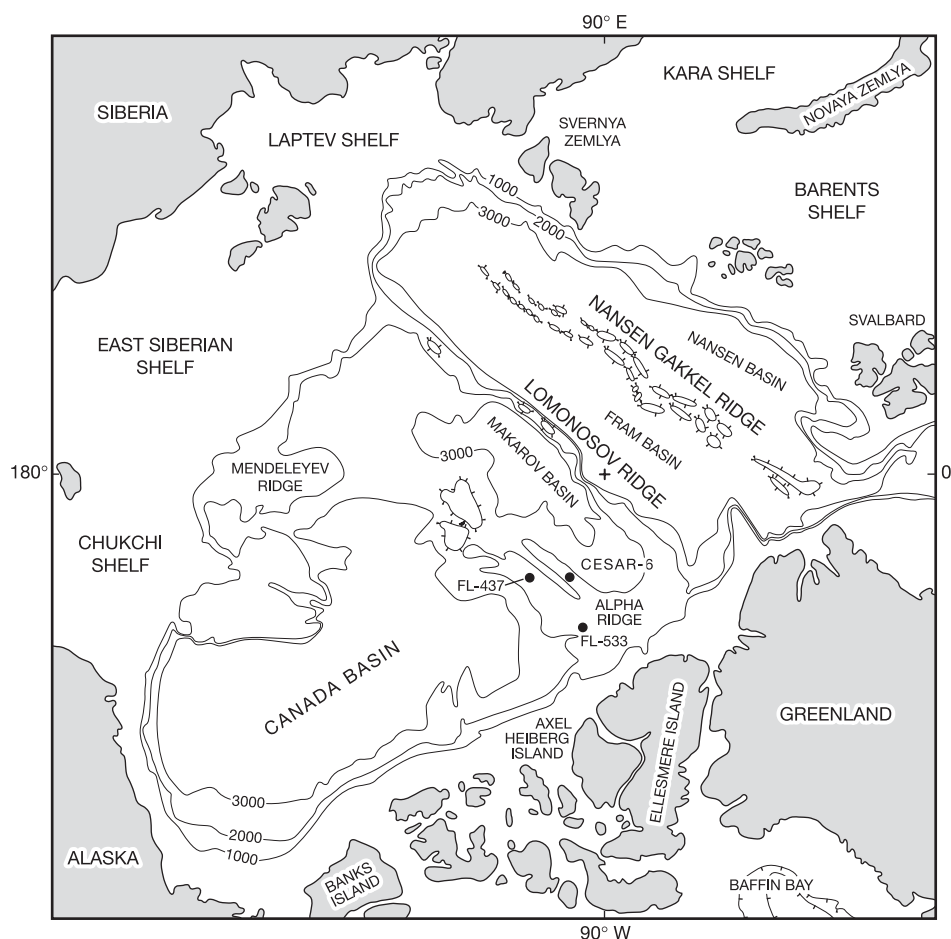


Figure 1 Map of the Arctic Ocean. Also shown (filled circles) are the locations of the three sites on the Alpha ridge where Cretaceous sediments have been retrieved. Core FL-533,

containing 67 cm of black mud, was cored at a water depth of 1,855 m on the eastern end of the structure (85° 05.9' N, 98° 17.8' W). Modified from ref. 9.

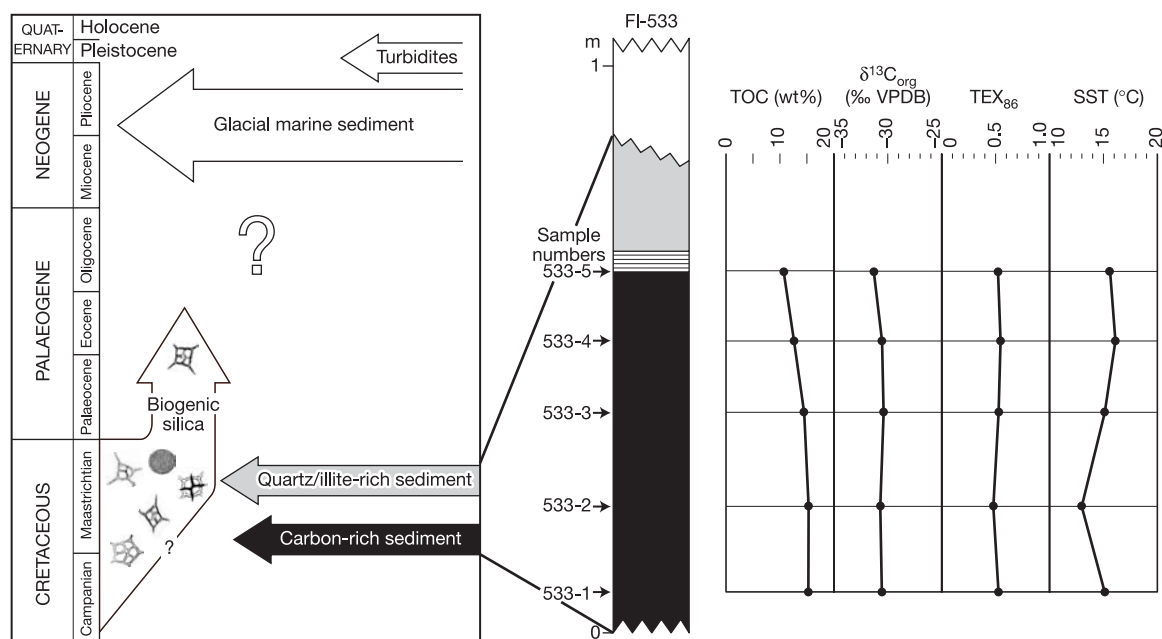


Figure 2 Suggested stratigraphic relationships of Campanian–Maastrichtian biogenic sediments of the Alpha ridge, Arctic Ocean. Our organic geochemical data from the black muds are shown against the lithological log for core FI-533. Stratigraphy modified from

ref. 1 to incorporate the more recent data of ref. 4. Sea surface temperature (SST), using the TEX_{86} parameter, calculated after the methodology given in refs 6, 16 and this Letter. TOC, total organic carbon.

Both sediments can be interpreted as the product of open-marine fertile waters, probably influenced by active upwelling centres that may have varied, with respect to their nutrient concentrations and planktonic biota, in time and space^{4,8,15}. Diatoms, however, were clearly a significant part of the biota when both siliceous and organic-rich sediments were accumulating, even though they have apparently left no obvious hard parts in the black muds. Strong seasonality is implied by the laminated varve-like nature of the siliceous sediments, with layers rich in resting spores of diatoms alternating with layers rich in their vegetative cells. The remarkably pristine state of preservation of the siliceous biota in cores FI-437 and CESAR-6 and the thermal immaturity of the marine organic matter (predominance (>80%) of $\beta\beta$ -homohopane isomers) in core FI-533 implies minimal diagenesis, presumably a function of the lack of sedimentary overburden. There are no dropstones in any of these sediments, so there is a complete lack of evidence for glacial activity.

Because the paucity of biogenic carbonate ($CaCO_3$ contents are <1% in the black muds²) in the sediments of all three cores precludes application of the oxygen-isotope palaeothermometer, we analysed five samples of the black shale from core FI-533 for membrane lipids of marine Crenarchaeota to determine the TEX_{86} ('tetraether index of 86 carbon atoms') parameter⁶. Marine Crenarchaeota are ubiquitous microorganisms that make up 20–30% of the picoplankton in present-day oceans, and their molecular fossils have been found in immature organic-rich sediments as old as Aptian (Early Cretaceous)¹⁶. Studies on marine core-top sediments have shown a significant ($r^2 = 0.92$) correlation between the number of cyclopentane rings in these lipids and annual mean sea surface temperature⁶. Furthermore, analysis of Cretaceous Atlantic black shales, using the TEX_{86} parameter, gave results that show excellent correlation with palaeotemperatures derived from $\delta^{18}O$ values of well-preserved glassy planktonic foraminifera¹⁶. The results of our analyses, shown in Fig. 2, indicate average Arctic Ocean ($\sim 80^\circ N$) sea surface temperatures of $15 \pm 1^\circ C$ for that part of the early Maastrichtian represented by the black muds.

The data presented here allow determination of the climatic gradient during the Late Cretaceous, commonly considered to be particularly subdued with respect to that of the present day^{17,18}. Oxygen-isotope data from Maastrichtian (probably early to mid-Maastrichtian) primary skeletal aragonite and enclosing magnesian-calcite cements deriving from a drowned carbonate platform (guyot) in the western Pacific indicate palaeo-equatorial (~ 5 – $10^\circ S$) temperatures of ~ 27 – $32^\circ C$ (ref. 19). Thus, assuming no rapid global climatic change during the early to mid-Maastrichtian—oxygen-isotope data from most sites in the world ocean imply a gradual surface-water cooling during this interval¹⁸—an Equator-to-pole gradient of $\sim 15^\circ C$ is implied for the Northern Hemisphere during this period of Late Cretaceous time. A remarkably similar figure of $\sim 14^\circ C$ difference between low- and high-latitude surface water was computed for the late Maastrichtian of the Southern Hemisphere, based on oxygen-isotope ratios of well-preserved planktonic foraminifera²⁰.

The Maastrichtian calibration point allows a tentative Mid- to Late Cretaceous palaeotemperature curve for the Arctic Ocean to be generated using $\delta^{18}O$ data from other localities. Relatively long $\delta^{18}O$ time series exist for southern England (Chalk), central Italy (Scaglia: lithified pelagic limestone), Ocean Drilling Program Sites on the Exmouth plateau off western Australia (chalks and claystones) and in the Southern Ocean where species-specific planktonic and benthonic foraminifera have been analysed^{21,22}. All these curves, whether or not derived from material that has been deeply buried and, in some cases, uplifted to experience possible meteoric-water diagenesis, show a remarkable similarity, demonstrating conclusively that the climatic trends are global in nature. The Cretaceous palaeotemperature profiles indicate, with some reversals, warming from the Albian through the Cenomanian, to peak either at Cenomanian–Turonian boundary time or in the early to middle Turonian, followed by a general decline through the Coniacian, Santonian and Campanian to reach a definitive minimum in the Maastrichtian. The smoothed curve from the English Chalk, taken to record broad temperature changes across the Northern Hemisphere, represents a composite record from outcrop

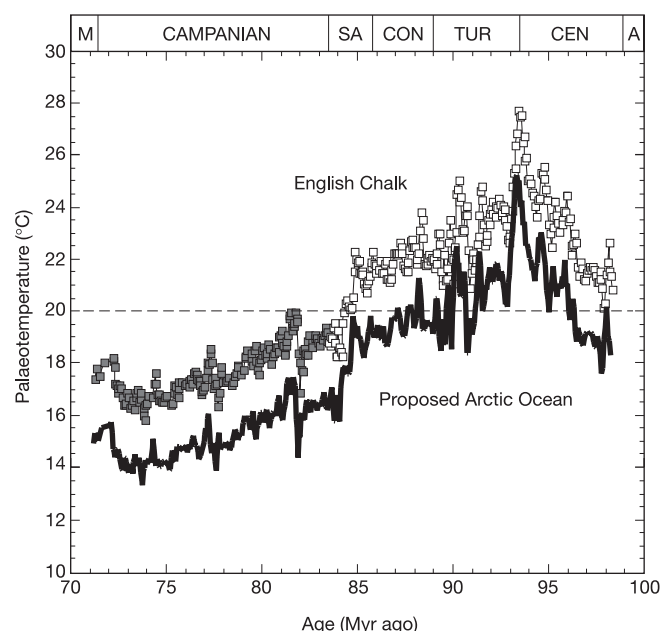


Figure 3 Proxy palaeotemperature curve for the Arctic Ocean. This curve (bold black line) is taken from oxygen-isotope ratios of bulk chalk from southern England (depositional palaeolatitude of $\sim 40^\circ\text{N}$; ref. 31), calibrated against a value of 15°C for the early Maastrichtian. Stratigraphic overlap in the lowest Campanian of the two curves (coastal outcrop in East Kent: unfilled squares) and borehole material in Norfolk (Trunch borehole: filled squares) permits adjustment of the $\delta^{18}\text{O}$ values, and hence calculated palaeotemperatures, of core samples. This adjustment is based on the assumption that the more lithified chalks record an additional burial diagenetic overprint: the core samples have consistently lower $\delta^{18}\text{O}$ values (by $\sim 0.7\text{‰}$) than coeval outcrop material. Palaeotemperatures on a smoothed (three-point moving average) curve are calculated using a standard formula, assuming a value of -1.0‰ SMOW for an ice-free Cretaceous ocean²¹. Note the 'tropical' polar climates (dashed line at 20°C) attained during the Cenomanian and Turonian stages.

(East Kent) and core material (Trunch borehole, Norfolk) joined at a level of the lowest Campanian (Fig. 3). Smoothing may reduce the significance of real episodes of short-period high and low temperatures but equally helps to eliminate the influence of diagenetic artefacts produced in intervals where post-depositional calcite cements dominate the isotopic signature.

Because the chalk palaeotemperature curve extends into the lowest Maastrichtian, it can be applied to the Arctic Ocean using the value of 15°C as a calibration point on the simplifying assumption that the latitudinal gradient did not greatly change over the time period in question. Using this proxy curve for the Arctic Ocean suggests average surface-water temperatures in excess of 20°C during much of the Cenomanian–Turonian interval (Fig. 3). There is palaeontological evidence to support the notion of such warm balmy climates at the North Pole (Fig. 4). Fossil leaves and pseudo-fruit of the breadfruit tree *Artocarpus dicksoni*, whose present-day 'ultra-tropical' relatives grow in a temperature range of $\sim 15\text{--}38^\circ\text{C}$ (ref. 23), were discovered as long ago as 1883 in Cenomanian sediments of western Greenland²⁴, and the skeletal remains of large crocodile-like champosaur from the Turonian of the Canadian Arctic archipelago (western Axel Heiberg island, Fig. 1) imply a mean annual ambient temperature exceeding 14°C (refs 25, 26). Interpretations from leaf physiognomy for the Turonian of the circum-Arctic region suggest somewhat lower temperatures²⁷. Similar palaeobotanical studies on Cretaceous plant-bearing units on the north slope of Alaska, taken to suggest mean annual air temperatures at sea level of 5°C in the Campanian–

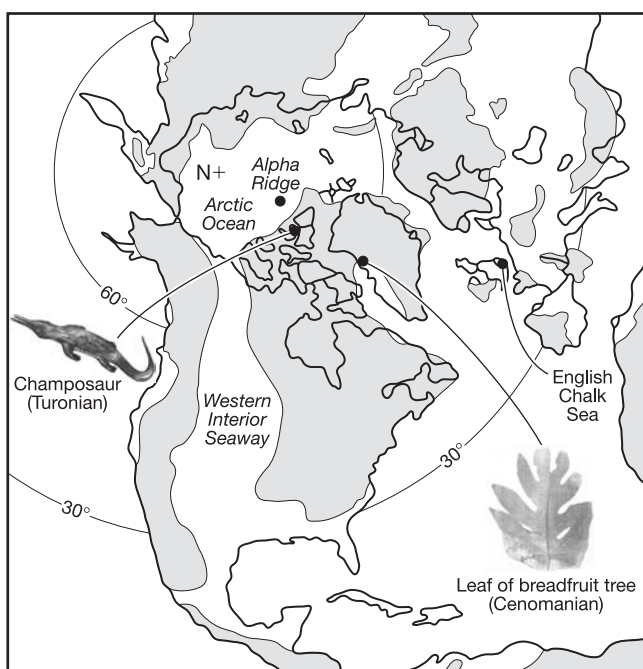


Figure 4 Northern Hemisphere view of the Arctic Ocean during the Late Cretaceous (Turonian). This projection shows the likely connections of the Arctic Ocean to epicontinental seas, such as that of the western interior of North America and the north European shelf³¹. Approximate location of English Chalk outcrops in southeast England also shown. Position of fossil remains of breadfruit-tree leaves (image of the actual specimen recovered shown) is after ref. 24. Position and image of champosaur are from refs 25 and 26, respectively. Approximate position of the Alpha ridge² indicates its relative proximity to land, as also suggested by the presence of abundant terrestrial material in the lower Maastrichtian black muds.

Maastrichtian²⁸, are incompatible with the results presented here, implying that all such physiognomic floral data need to be re-evaluated.

In conclusion, these warm Arctic marine palaeotemperatures are consistent with models of a CO_2 -rich Cretaceous atmosphere characterized by a vigorous hydrological cycle¹⁷. Exactly how a mean annual marine temperature of 15°C was resolved into seasonal variation is unknown, but this figure implies a total absence of polar ice at these high latitudes, and contrasts dramatically with a present-day mean annual surface air temperature of about -15°C at 80°N (ref. 17). □

Methods

Powdered and freeze-dried sediments (1–3 g dry mass) were extracted with dichloromethane (DCM)/methanol (2:1) by using the Dionex accelerated solvent extraction technique. The extracts were separated by Al_2O_3 column chromatography using hexane/DCM (9:1), DCM/methanol (95:5) and DCM/methanol (1:1) as subsequent eluents to yield the apolar, tetraether and polar fractions, respectively. The apolar and desulphurized (using Raney Ni) polar fractions were analysed by gas chromatography and gas chromatography/mass spectrometry. The tetraether fractions were analysed as in ref. 6 except that separation was achieved on a Prevail Cyano column ($2.1 \times 150\text{ mm}$, $3\text{ }\mu\text{m}$, with flow rate at 0.2 ml min^{-1}), and single ion monitoring of the $[\text{M} + \text{H}]^+$ ions (dwell time, 234 ms) was used to identify the tetraether lipids with 1–4 cyclopentane rings (see ref. 6) and calculate the TEX_{86} values. These values were converted to sea surface temperature (SST) according to the equation:

$$\text{TEX}_{86} = 0.015\text{SST} + 0.29 \quad (1)$$

Equation (1) is updated from ref. 6, and is based on correlating TEX_{86} values of >60 core-top sediments with annual mean SSTs ranging between 0 and 28°C . Replicate analysis has shown that the error in TEX_{86} values is ~ 0.01 or $\sim 1^\circ\text{C}$. This substantial improvement in analytical reproducibility (compare ref. 6) is mainly due to improved chromatographic conditions and the use of single-ion monitoring instead of a full scanning technique.

The high correlation of the TEX_{86} values with SST ($r^2 = 0.92$ for equation (1)) is in

agreement with the known physiological adaptation to temperature of the tetraether membrane lipid composition in cultured hyperthermophilic Archaea and their marine mesophilic descendants²⁹. Previous studies have also shown that the TEX₈₆ parameter is not sensitive to salinity or depositional redox conditions^{29,30}. Furthermore, initial applications show that reconstructed SSTs for Cretaceous sediments agree well with those derived from oxygen-isotope ratios of pristine skeletal carbonate¹⁶.

Bulk organic isotopes and TOC contents were determined by decalcifying powdered rock samples with 2 N hydrochloric acid and analysing the decalcified sediments in duplicate on a Carlo Erba 1112 Flash Elemental Analyser coupled to a ThermoFinnigan Delta Plus isotope mass spectrometer. Analytical errors for TOC range from 0.3% to 2%, and for $\delta^{13}\text{C}_{\text{org}}$ (‰ versus VPDB) are <0.1‰.

Received 10 June; accepted 28 October 2004; doi:10.1038/nature03143.

- Clark, D. L. Early history of the Arctic Ocean. *Paleoceanography* **3**, 539–550 (1988).
- Clark, D. L., Byers, C. W. & Pratt, L. M. Cretaceous black mud from the central Arctic Ocean. *Paleoceanography* **1**, 265–271 (1986).
- Mudie, P. J. & Blasko, S. M. in *Initial Geological Report on CESAR: The Canadian Expedition to Study the Alpha Ridge* (eds Jackson, H. R., Mudie, P. J. & Blasko, S. M.) 59–99 (Paper 84–22, Geol. Surv. Canada, Ottawa, 1985).
- Firth, J. V. & Clark, D. L. An early Maastrichtian organic-walled phytoplankton cyst assemblage from an organic-walled black mud in Core FI-533, Alpha Ridge: evidence for upwelling conditions in the Cretaceous Arctic Ocean. *Mar. Micropaleont.* **34**, 1–27 (1998).
- Jokat, W. Seismic investigations along the western sector of Alpha Ridge, Central Arctic Ocean. *Geophys. J. Int.* **152**, 185–201 (2003).
- Schouten, S., Hopmans, E. C., Schefuß, E. & Sinninghe Damsté, J. S. Distributional variations in marine crenarchaeotal membrane lipids: a new tool for reconstructing ancient sea water temperatures? *Earth Planet. Sci. Lett.* **204**, 265–274 (2002).
- Gradstein, F. M. et al. A Mesozoic time scale. *J. Geophys. Res.* **B 99**, 24051–24074 (1994).
- Dell'Agnese, D. J. & Clark, D. L. Siliceous microfossils from the warm Late Cretaceous and Early Cenozoic Arctic Ocean. *J. Paleontol.* **68**, 31–47 (1994).
- Stoffyn-Egli, P. Iron and manganese micro-precipitates with a Cretaceous biosiliceous ooze from the Arctic Ocean: possible hydrothermal source. *Geo-Mar. Lett.* **7**, 223–231 (1987).
- Mudie, P. J. in *Initial Geological Report on CESAR: The Canadian Expedition to Study the Alpha Ridge* (eds Jackson, H. R., Mudie, P. J. & Blasko, S. M.) 148–174 (Paper 84–22, Geol. Surv. Canada, Ottawa, 1985).
- Barron, J. A. in *Initial Geological Report on CESAR: The Canadian Expedition to Study the Alpha Ridge* (eds Jackson, H. R., Mudie, P. J. & Blasko, S. M.) 137–148 (Paper 84–22, Geol. Surv. Canada, Ottawa, 1985).
- Bukry, D. in *Initial Geological Report on CESAR: The Canadian Expedition to Study the Alpha Ridge* (eds Jackson, H. R., Mudie, P. J. & Blasko, S. M.) 125–135 (Paper 84–22, Geol. Surv. Canada, Ottawa, 1985).
- Sinninghe Damsté, J. S. et al. The rise of the rhizosolenid diatoms. *Science* **304**, 584–587 (2004).
- Holba, A. G. et al. Application of 24-norcholestanes for constraining source ages of petroleum. *Org. Geochem.* **29**, 1269–1283 (1998).
- Kitchell, J. A. & Clark, D. L. Late Cretaceous–Paleogene paleogeography and paleocirculation: evidence of north polar upwelling. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **40**, 135–165 (1982).
- Schouten, S. et al. Extremely high sea-surface temperatures at low latitudes during the middle Cretaceous as revealed by archaeal membrane lipids. *Geology* **31**, 1069–1072 (2003).
- Barron, E. J. A warm, equable Cretaceous: the nature of the problem. *Earth Sci. Rev.* **19**, 305–338 (1983).
- Barrera, E. & Savin, S. M. in *Evolution of the Cretaceous Ocean–Climate System* (eds Barrera, E. & Johnson, C. C.) 245–282 (Spec. Paper 332, Geol. Soc. Am., Boulder, Colorado, 1999).
- Wilson, P. A. & Opdyke, B. N. Equatorial sea-surface temperatures for the Maastrichtian revealed through remarkable preservation of metastable carbonate. *Geology* **24**, 555–558 (1996).
- Huber, B. T., Hodell, D. A. & Hamilton, C. P. Mid- to Late Cretaceous climate of the southern high latitudes. Stable isotopic evidence for minimal equator-to-pole thermal gradients. *Bull. Geol. Soc. Am.* **107**, 1164–1191 (1995).
- Jenkyns, H. C., Gale, A. S. & Corfield, R. M. Carbon- and oxygen-isotope stratigraphy of the English Chalk and Italian Scaglia and its palaeoclimatic significance. *Geol. Mag.* **131**, 1–34 (1994).
- Clarke, L. J. & Jenkyns, H. C. New oxygen-isotope evidence for long-term Cretaceous climate change in the Southern Hemisphere. *Geology* **27**, 699–702 (1999).
- Morton, J. F. *Fruits of Warm Climates* (Creative Resources Systems, Miami, 1987).
- Nathorst, A. G. Ueber die Reste eines Brotfruchtbaums ARTOCARPUS DICKSONI n. sp., aus den cenomanen Kriedeablagerungen Grönlands. *Kongl. Svenska Vetenskaps-Akad. Hand* **24**, 2–9 (1890).
- Tarduno, J. A. et al. Evidence for extreme climatic warmth from Late Cretaceous Arctic vertebrates. *Science* **282**, 2241–2244 (1998).
- Huber, B. T. Tropical paradise at the Cretaceous poles? *Science* **282**, 2199–2200 (1998).
- Herman, A. B. & Spicer, R. A. Palaeobotanical evidence for a warm Cretaceous Arctic Ocean. *Nature* **380**, 330–333 (1996).
- Spicer, R. A. & Parrish, J. T. Late Cretaceous–early Tertiary palaeoclimates of northern high latitudes: a quantitative view. *J. Geol. Soc. Lond.* **147**, 329–341 (1990).
- Wüchter, C., Schouten, S. & Sinninghe Damsté, J. S. Temperature-dependent variation in the distribution of tetraether membrane lipids of marine Crenarchaeota: Implications for TEX₈₆ paleothermometry. *Paleoceanography* (in the press).
- Schouten, S., Hopmans, E. & Sinninghe Damsté, J. S. The effect of maturity and depositional redox conditions on archaeal tetraether lipid palaeothermometry. *Org. Geochem.* **35**, 567–571 (2004).
- Hay, W. W., Eicher, D. L. & Diner, R. in *Evolution of the Western Interior Basin* (eds Caldwell, W. G. E. & Kauffman, E. G.) 297–318 (Spec. Pap. 39, Geol. Ass. Canada, St John's, Newfoundland, 1993).

Acknowledgements We thank D. Clark for indicating the whereabouts of core FI-533, and T. Simo for locating it in the Department of Geology and Geophysics of the University of Wisconsin at Madison. S. Rampen and J. Ossebaar (Royal NIOZ) are thanked for analytical assistance.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.C.J. (hughj@earth.ox.ac.uk).

Recycling lower continental crust in the North China craton

Shan Gao^{1,2}, Roberta L. Rudnick³, Hong-Ling Yuan¹, Xiao-Ming Liu¹, Yong-Sheng Liu², Wen-Liang Xu⁴, Wen-Li Ling², John Ayers⁵, Xuan-Che Wang² & Qing-Hai Wang⁴

¹Key Laboratory of Continental Dynamics, Department of Geology, Northwest University, Xi'an 710069, China

²Faculty of Earth Sciences, China University of Geosciences, Wuhan 430074, China

³Geochemistry Laboratory, Department of Geology, University of Maryland, College Park, Maryland 20742, USA

⁴School of Earth Sciences, Jilin University, Changchun 130061, China

⁵Department of Geology, Vanderbilt University, PO Box 105 Station B, Nashville, Tennessee 37235, USA

Foundering of mafic lower continental crust into underlying convecting mantle has been proposed as one means to explain the unusually evolved chemical composition of Earth's continental crust^{1,2}, yet direct evidence of this process has been scarce. Here we report that Late Jurassic high-magnesium andesites, dacites and adakites (siliceous lavas with high strontium and low heavy-rare-earth element and yttrium contents) from the North China craton have chemical and petrographic features consistent with their origin as partial melts of eclogite that subsequently interacted with mantle peridotite. Similar features observed in adakites and some Archaean sodium-rich granitoids of the tonalite-trondhjemite-granodiorite series have been interpreted to result from interaction of slab melts with the mantle wedge. Unlike their arc-related counterparts, however, the Chinese magmas carry inherited Archaean zircons and have neodymium and strontium isotopic compositions overlapping those of eclogite xenoliths derived from the lower crust of the North China craton. Such features cannot be produced by crustal assimilation of slab melts, given the high Mg#, nickel and chromium contents of the lavas. We infer that the Chinese lavas derive from ancient mafic lower crust that foundered into the convecting mantle and subsequently melted and interacted with peridotite. We suggest that lower crustal foundering occurred within the North China craton during the Late Jurassic, and thus provides constraints on the timing of lithosphere removal beneath the North China craton.

Eclogite forms by high- to ultrahigh-pressure metamorphism of basaltic rocks and has a density that is higher than that of peridotite by 0.2–0.4 g cm⁻³ (ref. 3). Because of this density contrast, mafic lower continental crust (together with the underlying lithospheric mantle) can be recycled into the mantle when granulite is transformed into eclogite during crustal thickening^{1,4}. Such foundering may explain collapse of mountains, basin formation and associated magmatism, which is generated by asthenospheric upwelling into the space previously occupied by thickened lithosphere⁵. Density foundering may also explain the absence of high-pressure mafic residues that are the complement to the voluminous tonalite-

trondhjemite-granodiorite (TTG) gneisses found in Archaean cratons.

The eastern block of the North China craton (Fig. 1) is perhaps the best example of an Archaean craton that has lost its lithospheric keel^{6–8} and is accordingly an ideal place to test the hypothesis of lower crustal foundering. Ancient remnants exposed at the surface or in lower crustal xenoliths^{9,10} demonstrate the presence of early Archaean crust (≥ 3.6 Gyr old), whereas the more dominant late Archaean and Palaeoproterozoic U–Pb ages record significant later reworking¹¹ (see below). The contrasting compositions of mantle xenoliths carried in ~ 460 Myr diamondiferous kimberlites from those in nearby Neogene alkaline basalts record removal of the Archaean lithospheric mantle during the Phanerozoic^{6–8}, most

probably during the Jurassic–Cretaceous^{8,12}, when the eastern block experienced significant reactivation, accompanied by extensive magmatism and large-scale basin formation. This voluminous magmatism, high surface heat flow¹³, slow seismic wave speeds and thin crust and lithosphere^{13,14} make the eastern block of the North China craton unique amongst Archaean cratons. This craton also has chemically evolved lower and bulk crust compositions¹³. The above lines of evidence have been used to infer that lower crustal recycling may have accompanied removal of the lithospheric mantle¹³.

Mesozoic high-magnesium (Mg) volcanic rocks from Xinglonggou erupted along a narrow belt in the eastern block of the North China craton (western Liaoning Province, Fig. 1). The 403-m-thick Xinglonggou Formation consists of high-Mg andesites, adakites ($\text{Na}_2\text{O}/\text{K}_2\text{O} > 2.0$) and dacites ($\text{Mg\#} = 53\text{--}65$, where $\text{Mg\#} = \text{molar } 100^* \text{Mg}/(\text{Mg} + \text{Fe})$), interlayered with their lower Mg# (38–42) equivalents and rhyolites (Table 1). Euhedral, needle-shaped igneous zircons in two Xinglonggou rhyolites from the lower section yield concordant $^{206}\text{Pb}/^{238}\text{U}$ ages of 159 ± 3 Myr (2 σ) (XL31) and 159 ± 4 Myr (XL34) (Fig. 2a; Supplementary Figs 1–3; Supplementary Table 1). Similar zircons from an aphanitic lava (XL03) from the upper section yields a less precise, younger age of 144 ± 9 Myr (Fig. 2a). Thus, the Xinglonggou lavas are Late Jurassic in age.

A striking feature of the Xinglonggou high-Mg adakites and dacites is the presence of octahedral chromite ($\text{Cr\#} = 65\text{--}81$ in the core) and zoned orthopyroxene phenocrysts, which have a mantle with distinctively high Mg, Cr and Ni (Fig. 3; Supplementary Table 4). The sharp and irregular boundary between the core and mantle indicates that the latter formed as an overgrowth, with little diffusive exchange occurring between the two regions. This zoning reflects a significant increase in melt Mg# by reaction between the melt and a high-Mg environment (for example, mantle peridotite), whereas the very thin, low-Mg# rims probably represent magmatic orthopyroxene crystallized at crustal depths.

The presence of chromite and orthopyroxene are reflected in the unusually high whole-rock Cr (≤ 400 p.p.m.) and Ni (≤ 310 p.p.m.) contents. Other striking geochemical features of most Xinglonggou lavas include high Na_2O (≤ 5.7 wt%) and Sr (500–1,618 p.p.m.), and depletion in heavy rare-earth elements (HREEs; $\text{Yb} < 1.8$ p.p.m.) and Y (≤ 18 p.p.m.), with high Sr/Y (36–135) and $\text{La}_\text{N}/\text{Yb}_\text{N}$ ratios (17–19; where subscript N denotes chondrite normalization) (Table 1). These features are characteristic of adakites from island arcs¹⁵ and TTG that dominate areas of Archaean crust¹⁶. Like these rocks, the Xinglonggou lavas also show Nb–Ta depletion relative to La, and Pb enrichment relative to Ce (Supplementary Table 5).

Zircon age populations provide further evidence for the origin of these lavas. Although rhyolites contain primarily Mesozoic igneous zircons (Fig. 2a), inherited zircons dominate in adakites, andesites and dacites (for example, XL03 and XL18). The latter show a prominent age cluster at 2,500 Myr, which coincides with the major crustal growth period of the North China craton¹¹ (Fig. 2a–c). This inheritance pattern is consistent with zircon saturation systematics¹⁷. Magmatic orthopyroxene in the Xinglonggou andesites and dacites suggest that their temperature was $> 800^\circ\text{C}$ (ref. 18), but zircon saturation temperatures are lower ($\sim 760^\circ\text{C}$, calculated for XL03 and XL18 having ~ 160 p.p.m. Zr)¹⁷. Thus, zircon is under-saturated in these melts and the observed crystals are xenocrysts, consistent with their old ages and sub-rounded morphologies (for example, XL18, Fig. 2b). Furthermore, if these melts formed from foundered lower crust, as we advocate below, they should be nearly anhydrous, which is consistent with the absence of hydrous phases in the Xinglonggou lavas. Harrison and Watson¹⁹ calculate that it could take > 100 Myr for a 50- μm -diameter zircon crystal to dissolve in an anhydrous granitic melt at 850°C , which is consistent with the persistence of xenocrysts in these lavas. In contrast,

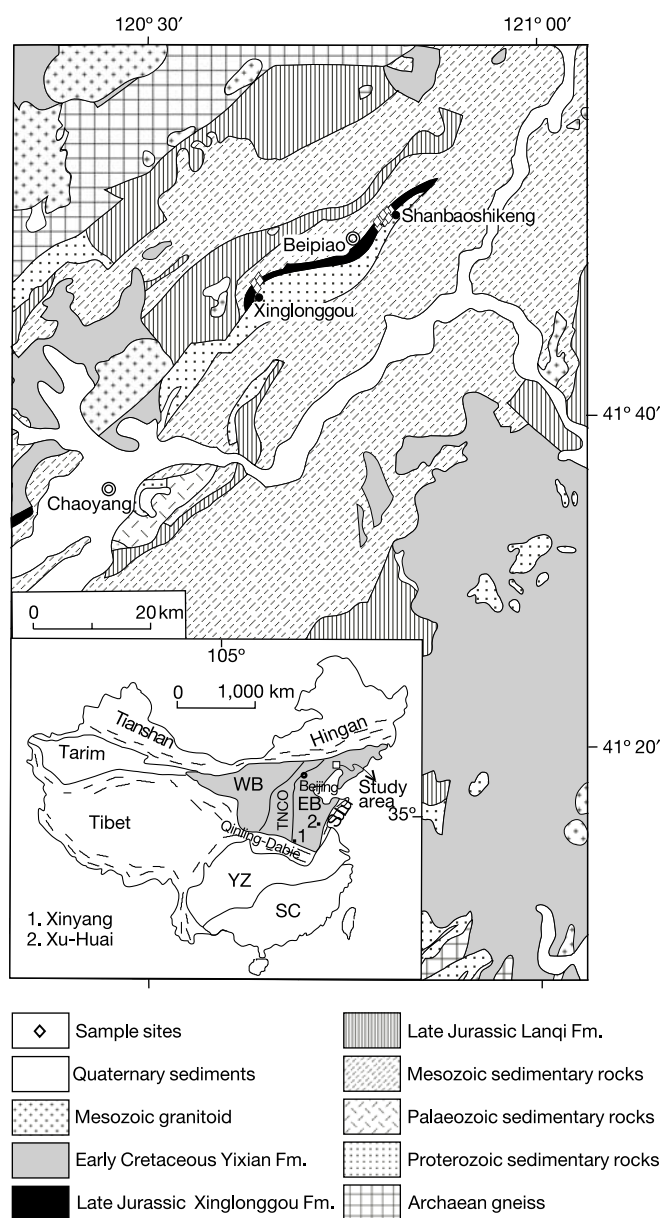


Figure 1 Geologic map of Western Liaoning Province, China. Our samples are from the Shanbaoshikeng and Xinglonggou type sections to the north and southwest of Beipiao, respectively. Inset shows major tectonic divisions of China, where YZ and SC denote the Yangtze craton and South China orogen, respectively. Also shown are the subdivisions of the North China craton¹¹, where WB, TNCO and EB denote the Western block, Trans-North China orogen and Eastern block, respectively. Numbers indicate the location of granulite, pyroxenite and eclogite xenoliths at (1) Xinyang and (2) Xu-Huai.

the rhyolites (XL31 and XL34) have significantly higher Zr (211–231 p.p.m.) and yield a zircon saturation temperature (820 °C) that is reasonable for these orthopyroxene-free magmas and consistent with the abundance of igneous zircons in the rhyolites.

Four origins have been proposed for adakites, high-Mg andesites and TTGs: (1) melting of subducted oceanic crust, followed by interaction with the overlying mantle wedge¹⁵; (2) melting of mantle peridotite under hydrous conditions²⁰; (3) melting of thickened mafic lower continental crust²¹, and (4) melting of foundered lower crust^{5,22}. In all of these models except for (2), depletion in HREEs and Y requires melting of a mafic source rock within the stability field of garnet, most probably under eclogite-facies conditions.

The geochemical and mineralogical characteristics of the Xinglonggou lavas can be used to rule out a number of the above hypotheses. These lavas cannot have originated by melting in the lower continental crust, as chromite will not crystallize from such melts and the increase in Ni, Cr and Mg# in orthopyroxene requires interaction with peridotite after the melt formed. Likewise, the lavas cannot have formed by direct melting of peridotite or by crustal contamination of an original picritic melt, as these scenarios would

predict high-Mg cores in the orthopyroxene, which is opposite to observation. Thus, the lavas may have originated as melts of either subducted oceanic crust or foundered lower crust with subsequent interaction of the melts with mantle peridotite. Indeed, the observed increase of Mg#, Ni and Cr of the melt (recorded in the orthopyroxene phenocrysts), together with orthopyroxene crystallization, are key features of slab melt-peridotite interaction^{15,23}.

The role played by Mesozoic subduction of the Palaeo-Pacific and Mongolo-Okhotsk oceanic plates in the genesis of the widespread volcanic and granitic magmatism along coastal eastern China (see, for example, ref. 24) is controversial. If a Mesozoic subduction zone did exist beneath eastern China, then the Xinglonggou lavas might have formed as slab melts. Adakites and associated high-Mg andesite from modern arcs that are interpreted as slab melts have mid-ocean-ridge basalt (MORB)-like Sr–Nd isotopic compositions, similar to those of the subducting oceanic crust (Fig. 4). In contrast, the Xinglonggou lavas have more evolved isotopic compositions (Fig. 4), with Nd model ages (T_{DM}) as old as 1.0 Gyr (Table 1). These isotopic signatures, together with the abundance of inherited Archaean zircons, requires that if the lavas originated by slab melting, they must have experienced significant crustal assimila-

Table 1 Geochemical and Sr–Nd isotopic compositions of selected Xinglonggou lavas

Sample	Lithology	Phenocrysts*	SiO ₂	Al ₂ O ₃	TFeO	MgO	CaO	Na ₂ O	K ₂ O	Mg#	Cr	Ni	Sr	Y	Zr	Yb	Sr/Y	La _N /Yb _N	Eu/Eu*
Geochemical compositions																			
XL01	HMAD		64.79	15.63	4.07	2.60	3.49	3.67	3.42	53	148	94	461	12	183	0.99	38	18	0.93
XL02	HMAD		63.11	15.12	3.98	3.54	5.15	3.56	1.24	61	137	85	716	9	163	0.98	80	17	0.98
XL03	HMAD	Opx†, Chm	63.09	15.25	4.02	3.56	5.28	3.35	1.30	61	149	89	765	11	162	1.1	70	17	0.84
XL04	HMAD	Opx†, Chm	62.59	15.27	4.08	3.64	5.47	3.28	1.16	61	147	86	836	13	162	1.1	64	17	0.84
XL05	HMAD	Opx†, Cpx, Chm	62.62	15.29	4.02	3.67	5.54	3.32	1.14	62	143	86	801	11	159	1.0	73	17	0.84
XL09	HMAD	Opx	64.86	15.26	3.77	2.93	3.54	3.78	3.37	58	127	82	491	13	161	1.1	38	18	0.84
XL12	HMAD	Chm	57.44	14.95	5.54	5.68	6.69	2.75	1.38	65	323	217	912	16	199	1.5	57	12	0.91
XL14	HMAD		57.35	15.71	7.02	4.47	5.46	3.34	2.37	53	313	246	626	19	201	1.9	33	10	0.90
XL15	AD	Ab	56.17	16.45	8.41	3.35	3.88	5.70	0.20	42	402	311	252	18	196	2.1	14	10	0.85
XL16	AN		59.15	16.51	6.88	2.33	2.51	5.36	3.38	38	344	310	613	17	210	1.8	36	11	0.96
XL17	DA		65.97	15.35	3.91	1.41	2.50	4.53	3.78	39	187	190	580	12	157	0.95	48	19	0.94
XL18	HMAD	Cpx	65.02	15.34	4.37	2.31	2.55	4.93	3.48	49	241	171	706	10	154	0.88	71	17	0.97
XL19	DA		66.00	15.48	3.99	1.63	1.99	5.14	3.75	42	231	166	729	10	162	0.88	73	18	0.98
XL20	HMAD	Cpx, Chm	62.94	15.00	5.23	2.60	2.70	3.87	3.55	47	337	105	617	15	154	1.4	41	11	0.94
XL21	HMAD		62.81	15.13	4.01	3.60	5.46	3.46	1.16	62	147	91	876	11	176	1.0	80	17	0.89
XL22	HMAD		62.22	15.46	4.15	3.81	5.59	3.16	1.25	62	137	92	1618	12	180	1.1	135	17	0.89
XL27	HMAD		62.05	15.22	4.11	3.72	5.40	3.39	1.23	62	130	90	763	12	174	1.1	64	17	0.87
XL31	RH		71.59	14.25	1.86	0.37	0.72	4.08	4.33	26	7.0	6.0	211	16	211	1.9	13	21	0.79
XL34	RH		70.92	14.96	2.08	0.35	0.83	5.26	3.52	23	22	41	231	21	231	2.0	11	19	0.71
Isotopic compositions																			
Sample	¹⁴³ Nd/ ¹⁴⁴ Nd		2σ	¹⁴⁷ Sm/ ¹⁴⁴ Nd		2σ	⁸⁶ Sr/ ⁸⁷ Sr		2σ	⁸⁷ Rb/ ⁸⁷ Sr		T_{DM}							
XL01	0.512489 ± 6			0.1091			0.707211 ± 4			0.7785		0.96							
XL02	0.512471 ± 7			0.1065			0.706470 ± 4			0.1252		0.97							
XL03	0.512511 ± 7			0.1074			0.705583 ± 15			0.1115		0.92							
XL04	0.512448 ± 7			0.1074			0.705772 ± 20			0.0868		1.01							
XL05	0.512511 ± 6			0.1076			0.705745 ± 16			0.0864		0.92							
XL09																			
XL12	0.512609 ± 7			0.1136			0.705338 ± 18			0.0671		0.82							
XL14																			
XL15	0.512667 ± 10			0.1203			0.705517 ± 17			0.0451		0.79							
XL16	0.512668 ± 6			0.1197			0.706390 ± 16			0.3584		0.78							
XL17	0.512575 ± 9			0.1101			0.706916 ± 16			0.5825		0.85							
XL18	0.512556 ± 5			0.1126			0.706417 ± 14			0.4092		0.90							
XL19	0.512525 ± 5			0.1116			0.706655 ± 13			0.4205		0.93							
XL20	0.512567 ± 9			0.1183			0.707304 ± 11			0.4916		0.93							
XL21	0.512413 ± 8			0.1074			0.705970 ± 13			0.0801		1.06							
XL22	0.512442 ± 4			0.1071			0.706955 ± 11			0.0332		1.01							
XL27	0.512517 ± 5			0.1070			0.706081 ± 12			0.1009		0.90							
XL31																			
XL34																			

Major oxides are reported in weight per cent (wt%), trace elements in parts per million (p.p.m.) and T_{DM} in billion years (Gyr). TFeO, total iron as FeO; Mg#, molar 100*Mg/(Mg + Fe). Eu/Eu* = $Eu_N / (Sm_N \times Gd_N)^{1/2}$, where subscript N denotes chondrite normalization. AD, adakite; AN, andesite; DA, dacite; HMAD, high-Mg adakite; HMAN, high-Mg andesite; HMDA, high-Mg dacite; RH, rhyolite. High-Mg rocks are defined as those with Mg# > 45. Ab, albite; Chm, chromite; Cpx, clinopyroxene; Opx, orthopyroxene. The Nd model age based on depleted mantle (T_{DM}) assumes a linear evolution of isotopic composition from $\epsilon_{Nd}(T) = 0$ at 4.56 Gyr to approximately +10 at the present time, and was calculated using the equation: $T_{DM} = (1/\lambda) \ln[1 + ((^{143}\text{Nd}/^{144}\text{Nd})_{\text{sample}} - 0.51315)/((^{147}\text{Sm}/^{144}\text{Nd})_{\text{sample}} - 0.2137)]$, where the decay constant (λ) of ¹⁴⁷Sm used in the model age calculation is 0.00654 Gyr⁻¹.

* Determined by electron microprobe.

† Opx shows a contrasting lower-Mg core versus a higher-Mg mantle.

tion. However, there are no correlations between SiO_2 and initial Sr isotopic compositions at 130 or 160 Myr (correlation coefficient $r = 0.19\text{--}0.32$), which is inconsistent with the evolved isotopic compositions resulting from assimilation and fractional crystallization (AFC)-like processes in the lower crust (see Fig. 4 for models). Similarly, there is no correlation ($r = -0.11$ to 0.12) between initial $^{87}\text{Sr}/^{86}\text{Sr}$ and Rb contents in the Xinglonggou lavas, as would be expected if the high $^{87}\text{Sr}/^{86}\text{Sr}$ results from assimilation of high-Rb/Sr crust. The Xinglonggou high-Mg adakites, andesites and dacites do show positive correlations between initial $^{143}\text{Nd}/^{144}\text{Nd}$ and Cr and Ni ($r = 0.85\text{--}0.89$), which can be generated by interaction of the Xinglonggou adakitic melts with peridotite, whereby transition metal concentrations in the melt increase owing to consumption of olivine.

We propose that the Xinglonggou lavas formed by melting of foundered eclogite that formed at the base of Archaean lower crust that was thickened in the Triassic. The adakitic melts thus produced interacted with peridotite upon ascent to produce the high Mg, Cr

and Ni signatures, the chromites and the zoned orthopyroxene phenocrysts.

Eclogite and garnet-clinopyroxenite xenoliths in Mesozoic volcanoclastic diatremes at Xinyang²⁵ and Mesozoic dioritic-monzodioritic porphyries at Xu-Huai²⁶ (Fig. 1) document the existence of eclogite within the deep (1.3–1.7 GPa) lithosphere of the North China craton during the Mesozoic. The Xu-Huai eclogite yields a whole rock-garnet Sm–Nd isochron of 219 ± 5 Myr (ref. 26), which overlaps with the 220–240 Myr eclogite-facies metamorphism in the Dabie-Sulu ultrahigh-pressure metamorphic belt (ref. 27 and references therein). U–Pb dates of the few available zircons from the Xu-Huai xenoliths reveal inheritance ages of 2.3–2.5 and 1.7–2.0 Gyr (Supplementary Table 6). Thus the eclogites formed in the Triassic from pre-existing Archaean lower crust, are not related to Jurassic-Cretaceous subduction of Palaeo-Pacific and Mongolo-Okhotsk oceanic crusts, and could be a plausible source for the adakitic melts. Indeed, modelling of the Sr–Nd isotopic compositions (Fig. 4) and Ni–Cr concentrations (not shown) of the

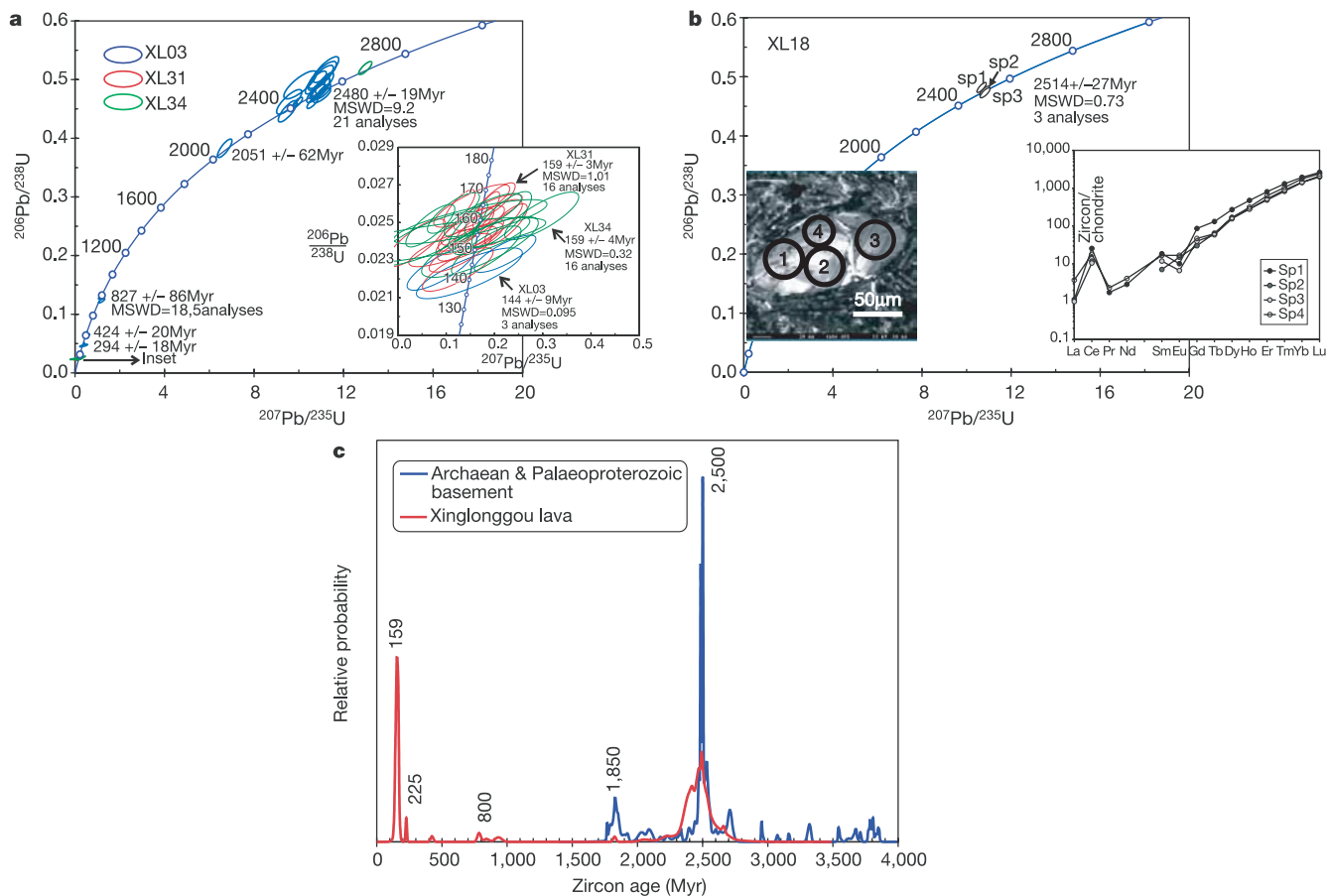


Figure 2 Zircon ages of the Xinglonggou lavas compared to those of the North China Archaean and Palaeoproterozoic basement. Concordia plots for zircons from **a**, high-Mg adakite (XL03) and rhyolites (XL31 and XL34), and **b**, high-Mg dacite (XL18). Data for XL03, XL31 and XL34 were determined by SHRIMP II. Inset in **a** shows concordia plot for euhedral, needle-shaped zircons of volcanic origin in the rhyolites (Supplementary Figs 1–3; Supplementary Table 1). Igneous zircons from the overlying lava, XL03, are significantly finer-grained ($<50\text{ }\mu\text{m}$ in diameter), consistent with its aphanitic texture, and yield a slightly younger weighted mean $^{206}\text{Pb}/^{238}\text{U}$ age of 144 ± 9 Myr. Three $40\text{-}\mu\text{m}$ spots (Sp1, 2 and 3) of a large ($75 \times 163\text{ }\mu\text{m}$), sub-rounded, inherited zircon from XL18 were measured in thin section for both the U–Pb age and REE concentrations by excimer LA-ICP-MS (Supplementary Tables 2, 3). An additional $30\text{-}\mu\text{m}$ spot (Sp4) was analysed only for REE composition. Left inset in **b** shows cathodoluminescence image of this zircon

with spot locations marked. Right inset shows chondrite-normalized REE distributions of the four spots. Error ellipses are shown at 1σ . Ages with MSWD are the weighted mean $^{207}\text{Pb}/^{206}\text{Pb}$ age of zircons older than 1,000 Myr and weighted mean $^{206}\text{Pb}/^{238}\text{U}$ ages for zircons younger than 1,000 Myr. Uncertainties in ages are quoted at the 95% confidence level (2σ). Note that the Archaean ages for XL03 and XL18 are within error of each other. **c**, Comparison of zircon age distribution between Xinglonggou lavas (red) and North China Archaean and Palaeoproterozoic basement (blue). Ages for the Archaean and Palaeoproterozoic basement are from the literature, and were determined by SHRIMP and the single grain evaporation $^{207}\text{Pb}/^{206}\text{Pb}$ method (see Supplementary Data). The $^{207}\text{Pb}/^{206}\text{Pb}$ age is used for zircons older than 1,000 Myr, while the $^{206}\text{Pb}/^{238}\text{U}$ age is used for younger ones.

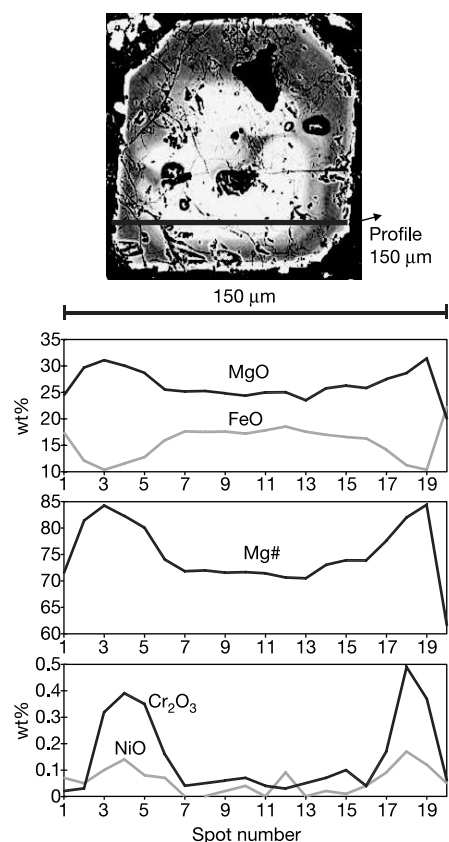


Figure 3 Compositional variations in an euhedral orthopyroxene phenocryst along the [001] crystallographic plane from Xinglonggou high-Mg adakite XL03. In the backscattered electron image (top) the dark areas are Mg-rich and the light areas are Fe-rich. In contrast to the grain's core, its mantle shows higher MgO, Cr and Ni contents, which reduce sharply in the very thin rim. The sharp and irregular boundary between the grain's core and mantle indicates that the mantle is a later overgrowth by chemical reaction, with little diffusive exchange between the two regions. In contrast, the mantle-rim boundary is regular.

Xinglonggou lavas suggests their derivation from a source region similar to the Xu-Huai eclogites and garnet clinopyroxenites, with subsequent assimilation (at <40% magma consumption) of peridotite having an isotopic composition of MORB.

As dynamical models show⁴, density foundering of lower crust is likely to be accompanied by removal of lithospheric mantle. The latter was previously documented in the eastern block of the North China craton^{6–8}, and probably occurred during the Late Jurassic or Early Cretaceous^{8,12}. Our data indicate that foundering was initiated by 159 Myr ago and probably was protracted (lasting into the Early Cretaceous). However, the mechanism triggering the density instability and foundering is still an open question. It may have been related to subduction of the Palaeo-Pacific Ocean, to collision of an amalgamated North China-Mongolian plate with the Siberian plate during closure of a Mongolo-Okhotsk ocean²⁴, or to more global events, like mantle overturn¹²; it is also possible that a combination of all three of these mechanisms triggered the density instability, as they may all have operated during the Jurassic-Cretaceous. Upwelling of hot asthenosphere at the base of the crust followed lithospheric removal, and caused the intense tectonic reactivation and crustal melting that characterizes the North China craton, with production of highly fractionated granites (which host the most important Au–Cu and rare metal mineralization in eastern China) between 120 and 130 Myr ago²⁸. This was followed by basin formation and eruption of asthenosphere-derived alkali basalts that started 100 Myr ago and lasted to the Holocene⁷. All these

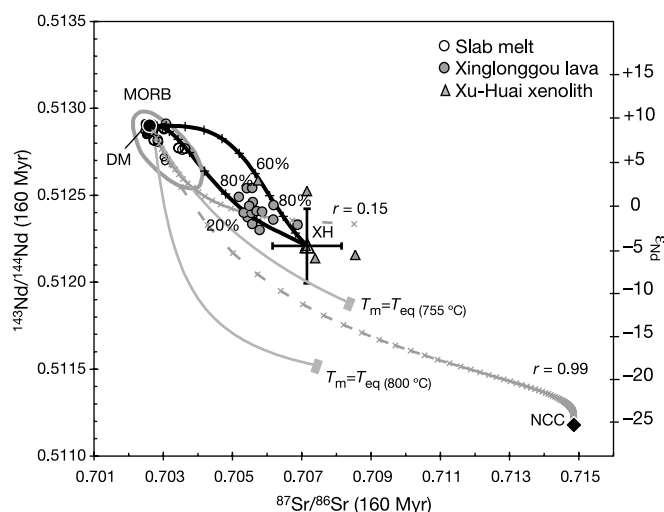


Figure 4 Sr–Nd isotopic compositions of the Xinglonggou lavas (this work, calculated at 160 Myr, the time of volcanic eruption) and inferred modern slab melts (open circles; small open circles indicate associated basalts). The thick black curves with pluses are AFC (assimilation and fractional crystallization) trends showing 10% increments in F (magma remaining) for a 30% partial melt of the median Xu-Huai eclogite-garnet clinopyroxenite (XH; large open triangle with 1σ error bars) that assimilates mantle (DM) with the isotopic composition of mid-ocean ridge basalt (MORB). The two trends represent melts derived from xenoliths with the highest (upper curve) and lowest (lower curve) Sr–Nd contents, respectively. Also shown are AFC and energy-constrained AFC (EC-AFC) models for assimilation of slab melt by the North China crust (NCC). The AFC trend (grey broken thin curve with crosses indicating 5% increment in F) passes through the Xinglonggou data only for a very low rate of assimilation ($r \leq 0.15$), with less than 20% melt left after assimilation. Using the highest Cr (636 p.p.m.) and Ni (132 p.p.m.) contents in Aleutian high-Mg andesites/adakites as representative of adakite before continental crust assimilation, the modelled Cr and Ni in the corresponding melt is <2 p.p.m., which is far lower than observed in the Xinglonggou adakites and andesites (Table 1). EC-AFC trends (grey thin curves) for equilibration temperatures (T_{eq}) of 800 °C and 755 °C are shown. The EC-AFC trend passes through the Xinglonggou data only when the equilibration temperature is very close to assimilant solidus temperature (that is, <4 °C difference) (not shown for clarity). Such temperatures are considerably lower than the temperature (>800 °C) suggested by the presence of orthopyroxene phenocrysts and zircon saturation thermometry (as discussed in the main text). The above modelling indicates that the Xinglonggou lavas probably formed by consumption of <40% melt during reaction with the mantle, and did not originate from crustal assimilation of a slab melt. See Supplementary Discussion for a detailed explanation, data sources and modelling parameters.

events are consequences of lower crustal and lithospheric mantle foundering^{1,2,4,5}. □

Methods

U–Pb dating

Zircons from the volcanic rocks were dated on a SHRIMP II at the Beijing SHRIMP Center and on an excimer (193 nm wavelength) laser ablation inductively coupled plasma mass spectrometer (LA-ICP-MS) at the Key Laboratory of Continental Dynamics, Northwest University. The ICP-MS used is an Elan 6100 DRC (Dynamic Reaction Cell) from Perkin Elmer/SCIEX (Canada). The GeoLas 200M laser-ablation system (MicroLas, Göttingen, Germany) was used for the laser ablation experiments. Backscatter electron and cathodoluminescence images (Supplementary Figs 1–3) were made of the zircons before analyses. Uncertainties in ages are quoted at the 95% confidence level (2σ). Spot diameter was 30 and 30–40 μm for SHRIMP II and LA-ICP-MS, respectively. Common Pb corrections were made using measured ²⁰⁴Pb (Supplementary Tables 1, 6).

The SHRIMP II analyses followed established methods²⁹. The calibration standard is Sri Lankan gem zircon standard (SL13), and the internal standard is the Australian National University zircon standard TEMORA 1 (ref. 30).

For LA-ICP-MS analysis, raw count rates were measured for ²⁹Si, ²⁰⁴Pb, ²⁰⁶Pb, ²⁰⁷Pb, ²⁰⁸Pb, ²³²Th and ²³⁸U. U, Th and Pb concentrations were calibrated by using ²⁹Si as an internal standard and NIST SRM 610 as the reference standard. ²⁰²Hg is usually <10 counts per second in the gas blank. Therefore, the contribution of ²⁰⁴Hg to ²⁰⁴Pb was negligible and no correction was made. ²⁰⁷Pb/²⁰⁶Pb, ²⁰⁶Pb/²³⁸U and ²⁰⁸Pb/²³²Th ratios

were calculated using GLITTER 4.0 (Macquarie University), which were then corrected for both instrumental mass bias and depth-dependent elemental and isotopic fractionation using Harvard zircon 91500 as external standard. The ages were calculated using ISOPLOT³¹. Our measurement of TEMORA 1 as an unknown yielded a weighted ²⁰⁶Pb/²³⁸U age of 415 ± 4 Myr (MSWD (mean squared weighted deviates) = 0.112, $n = 24$)³², which is in good agreement with the recommended isotope dilution-thermal ionization mass spectrometry (ID-TIMS) age of 416.75 ± 0.24 Myr (ref. 30). Analytical details for age and trace and REE determinations of zircons are reported elsewhere³².

Mineral analysis

Major element composition of minerals was analysed on a JEOL Superprobe JXA 8100 at the Department of Geology, Peking University, using an accelerating voltage of 15 kV, beam current 1 × 10⁻⁸ A and spot diameter of 1 µm.

Major and trace element analysis

Chemical compositions of whole rocks were measured at the Key Laboratory of Continental Dynamics, Northwest University. Major element compositions were analysed by X-ray fluorescence (XRF; Rikagu RIX 2100) using fused glass disks. Trace element composition was analysed by ICP-MS (Elan 6100 DRC) after acid digestion of samples in Teflon bombs. They were also analysed by XRF using powdered rock pellets. Concentrations of Sr, Y, Nb, Zr, Cr and Ni obtained using these two methods for the same samples generally agree to within 10% uncertainty³³. Analyses of United States Geological Survey rock standards (BCR-2, BHVO-1 and AGV-1) indicate precision and accuracy better than 5% for major elements and 10% for trace elements and REEs³³.

Sr-Nd isotopes

Sr-Nd isotopic compositions were determined on a Finnigan MAT-261 mass spectrometer operated in static mode at the Isotope Laboratory of China, University of Geosciences, Wuhan. Two aliquots of sample powder (200 mesh), ~100 mg each, were weighed with one aliquot being added to a mixed solution of ⁸⁴Sr, ⁸⁵Rb, ¹⁴⁵Nd and ¹⁴⁹Sm isotope spikes. Samples were digested in Teflon bombs with a mixture of concentrated HF, HNO₃ and HClO₄. Columns of AG50W-X8 and HDEHP resins were used sequentially for separation and purification of REEs and finally for separation of Nd and Sm by HCl eluents. The measured ¹⁴³Nd/¹⁴⁴Nd and ⁸⁷Sr/⁸⁶Sr ratios were normalized to ¹⁴⁶Nd/¹⁴⁴Nd = 0.721900 and ⁸⁸Sr/⁸⁶Sr = 8.375209, respectively. The La Jolla standard measured during the course of analyses gave an average ¹⁴³Nd/¹⁴⁴Nd of 0.511862 ± 5 (2σ, $n = 15$), and BCR-2 gives ¹⁴³Nd/¹⁴⁴Nd = 0.512635 ± 4, ¹⁴⁷Sm/¹⁴⁴Nd = 0.1369, Nd = 29.10 p.p.m. and Sm = 6.591 p.p.m. NBS-987 gave ⁸⁷Sr/⁸⁶Sr = 0.710236 ± 16 (2σ, $n = 6$).

Received 27 May; accepted 3 November 2004; doi:10.1038/nature03162.

1. Kay, R. W. & Kay, S. M. Creation and destruction of lower continental crust. *Geol. Rundsch.* **80**, 259–278 (1991).
2. Rudnick, R. L. Making continental crust. *Nature* **378**, 571–578 (1995).
3. Rudnick, R. L. & Fountain, D. M. Nature and composition of the continental crust: a lower crustal perspective. *Rev. Geophys.* **33**, 267–309 (1995).
4. Jull, M. & Kelemen, P. B. On the conditions for lower crustal convective instability. *J. Geophys. Res.* **106**, 6423–6446 (2001).
5. Kay, R. W. & Kay, S. M. Delamination and delamination magmatism. *Tectonophysics* **219**, 177–189 (1993).
6. Menzies, A., Fan, W.-M. & Zhang, M. in *Magmatic Processes and Plate Tectonics* (eds Prichard, H. M., Alabaster, H. M., Harris, T. & Neary, C. R.) 71–81 (Geol. Soc., London, 1993).
7. Griffin, W. L., Zhang, A. D., O'Reilly, S. Y. & Ryan, C. G. in *Mantle Dynamics and Plate Interactions in East Asia* (eds Flower, M., Chung, S.-L., Lo, C.-H. & Lee, T.-Y.) 107–126 (American Geophysical Union, Washington DC, 1998).
8. Gao, S., Rudnick, R. L., Carlson, R. W., McDonough, W. F. & Liu, Y. S. Re-Os evidence for replacement of ancient mantle lithosphere beneath the North China Craton. *Earth Planet. Sci. Lett.* **198**, 307–322 (2002).
9. Liu, D.-Y., Nutman, A. P., Compston, W., Wu, J. S. & Shen, Q.-H. Remnants of > 3800 Ma crust in the Chinese part of the Sino-Korean craton. *Geology* **20**, 339–342 (1992).
10. Zheng, J. P. *et al.* 3.6 Ga lower crust in central China: New evidence on the assembly of the North China Craton. *Geology* **32**, 229–232 (2004).
11. Zhao, G. C., Wilde, S. A., Cawood, P. A. & Sun, M. Archean blocks and their boundaries in the North China Craton: lithological, geochemical, structural and P-T path constraints and tectonic evolution. *Precamb. Res.* **107**, 45–73 (2001).
12. Wilde, S. A., Zhou, X. H., Nemchin, A. A. & Sun, M. Mesozoic crust-mantle interaction beneath the North China craton: A consequence of the dispersal of Gondwanaland and accretion of Asia. *Geology* **31**, 817–820 (2003).
13. Gao, S. *et al.* How mafic is the lower continental crust? *Earth Planet. Sci. Lett.* **106**, 101–117 (1998).
14. Huang, Z. X., Su, W., Peng, Y. J., Zheng, Y. J. & Li, H. Y. Rayleigh wave tomography of China and adjacent regions. *J. Geophys. Res.* **108**, doi:10.1029/2001JB001696 (2003).
15. Kay, R. W. Aleutian magnesian andesites – melts from subducted Pacific ocean crust. *J. Volcanol. Geotherm. Res.* **4**, 117–132 (1978).
16. Martin, H. Adakitic magmas: modern analogues of Archaean granitoids. *Lithos* **46**, 411–429 (1999).
17. Watson, E. B. & Harrison, T. M. Zircon saturation revisited: Temperature and composition effects in a variety of crustal magma types. *Earth Planet. Sci. Lett.* **64**, 295–304 (1983).
18. Wolf, M. B. & Wyllie, P. J. Dehydration-melting of amphibolite at 10 kbar: the effects of temperature and time. *Contrib. Mineral. Petrol.* **115**, 369–383 (1994).
19. Harrison, T. M. & Watson, E. B. Kinetics of zircon dissolution and zirconium diffusion in granitic melts of variable water content. *Contrib. Mineral. Petrol.* **84**, 66–72 (1983).
20. Stern, R. A. & Hanson, G. N. Archean high-Mg granodiorite: a derivative of light rare earth element-enriched monzodiorite of mantle origin. *J. Petrol.* **32**, 201–238 (1991).
21. Atherton, M. P. & Petford, N. Generation of sodium-rich magmas from newly underplated basaltic crust. *Nature* **362**, 144–146 (1993).

22. Xu, J. F., Shinjo, R., Defant, M. J., Wang, Q. A. & Rapp, R. P. Origin of Mesozoic adakitic intrusive rocks in the Ningzhen area of east China: Partial melting of delaminated lower continental crust? *Geology* **30**, 1111–1114 (2002).
23. Rapp, R. P., Shimizu, N., Norman, M. D. & Applegate, G. S. Reaction between slab-derived melts and peridotite in the mantle wedge: experimental constraints at 3.8 GPa. *Chem. Geol.* **160**, 335–356 (1999).
24. Davis, G. A., *et al.* in *Paleozoic and Mesozoic Tectonic Evolution of Central Asia: From Continental Assembly to Intracratinal Deformation* (eds Hendrix, M. S. & Davis, G. A.) 171–197 (Memoir Geol. Soc. Am., Boulder, 2001).
25. Zheng, J. P., Sun, M., Lu, F. X. & Pearson, N. Mesozoic lower crustal xenoliths and their significance in lithospheric evolution beneath the Sino-Korean Craton. *Tectonophysics* **361**, 37–60 (2003).
26. Xu, W. L., Wang, D. Y., Liu, X. C., Wang, Q. H. & Lin, J. Q. Discovery of eclogite inclusions and its geological significance in early Jurassic intrusive complex in Xuzhou, northern Anhui, eastern China. *Chin. Sci. Bull.* **47**, 1212–1216 (2002).
27. Ayers, J. C., Dunkle, S., Gao, S. & Miller, C. F. Constraints on timing of peak and retrograde metamorphism in the Dabie Shan Ultrahigh-Pressure Metamorphic Belt, east-central China, using U–Th–Pb dating of zircon and monazite. *Chem. Geol.* **186**, 315–331 (2002).
28. Qiu, Y. M., Groves, D. I., McNaughton, N. J., Wang, L. G. & Zhou, T. H. Nature, age, and tectonic setting of granitoid-hosted, orogenic gold deposits of the Jiaodong Peninsula, eastern North China craton, China. *Mineralium Deposita* **37**, 283–305 (2002).
29. Williams, I. S. in *Applications of Microanalytical Techniques to Understanding Mineralizing Processes* (eds McKibben, M. A., Shanks, W. C. III & Ridley, W. L.) 1–35 (Reviews in Economic Geology, Vol. 7, Society of Economic Geologists, Littleton, Colorado, 1998).
30. Black, L. P. *et al.* TEMORA 1: a new zircon standard for Phanerozoic U–Pb geochronology. *Chem. Geol.* **200**, 155–170 (2003).
31. Ludwig, K. R. *ISOPLOT 3.00: A Geochronological Toolkit for Microsoft Excel* (Berkeley Geochronology Center, Berkeley, California, 2003).
32. Yuan, H. L. *et al.* Accurate U–Pb age and trace element determination of zircon by laser ablation-inductively coupled plasma mass spectrometry. *Geostand. Newsl.* (in the press).
33. Rudnick, R. L., Gao, S., Ling, W. L., Liu, Y. S. & McDonough, W. F. Petrology and geochemistry of spinel peridotite xenoliths from Hannuoba and Qixia, North China craton. *Lithos* **77**, 609–637 (2004).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This research was supported by the National Natural Science Foundation of China, Chinese Ministry of Science and Technology and the Beijing SHRIMP Center (S.G. and H.L.Y.) and the NSF (R.L.R. and J.C.A.). We thank D. Y. Liu, Z. C. Hu, G. M. Shu and Y. B. Wang for help in SHRIMP II and electron microprobe analyses, and D. Günther for help in setting up the LA-ICP-MS. This Letter benefited from comments and suggestions from S. Wilde and H. Rollinson.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.L.R. (rudnick@geol.umd.edu) or S.G. (sgao@263.net.cn).

The impact of surface-adsorbed phosphorus on phytoplankton Redfield stoichiometry

Sergio A. Sañudo-Wilhelmy¹, Antonio Tovar-Sanchez^{1,2}, Fei-Xue Fu³, Douglas G. Capone⁴, Edward J. Carpenter⁵ & David A. Hutchins³

¹Marine Sciences Research Center, Stony Brook University, Stony Brook, New York 11794-5000, USA

²IMEDEA (CSIC-UIB), Instituto Mediterráneo de Estudios Avanzados, Esporles 07190, Mallorca, Islas Baleares, Spain

³College of Marine Studies, University of Delaware, 700 Pilottown Road, Lewes, Delaware 19958, USA

⁴Wrigley Institute of Environmental Studies and Department of Biological Sciences, University of Southern California, Los Angeles, California 90089, USA

⁵Romberg Tiburon Center, San Francisco State University, 3152 Paradise Drive, Tiburon, California 94920, USA

The Redfield ratio of 106 carbon:16 nitrogen:1 phosphorus in marine phytoplankton¹ is one of the foundations of ocean biogeochemistry, with applications in algal physiology², palaeoclimatology³ and global climate change⁴. However, this ratio varies substantially in response to changes in algal nutrient status⁵ and taxonomic affiliation^{6,7}. Here we report that Redfield ratios are also strongly affected by partitioning into surface-adsorbed

and intracellular phosphorus pools. The C:N:surface-adsorbed P (80–105 C:15–18 N:1 P) and total (71–80 C:13–14 N:1 P) ratios in natural populations and cultures of *Trichodesmium* were close to Redfield values and not significantly different from each other. In contrast, intracellular ratios consistently exceeded the Redfield ratio (316–434 C:59–83 N:1 intracellular P). These high intracellular ratios were associated with reduced N_2 fixation rates, suggestive of phosphorus deficiency. Other algal species also have substantial surface-adsorbed phosphorus pools, suggesting that our *Trichodesmium* results are generally applicable to all phytoplankton. Measurements of the distinct phytoplankton phosphorus pools may be required to assess nutrient limitation accurately from elemental composition. Deviations from Redfield stoichiometry may be attributable to surface adsorption of phosphorus rather than to biological processes, and this scavenging could affect the interpretation of marine nutrient inventories and ecosystem models.

We collected colonies of *Trichodesmium* spp. using trace-metal-clean methods in the subtropical western Atlantic Ocean in April 2003 (Supplementary Table 1), and analysed them for surface-adsorbed and intracellular P using the oxalate surface wash method⁸. We also measured total and surface-adsorbed C, N, Fe, Mn and Mo. Additional laboratory experiments determined the removal efficiency of the surface-adsorbed P, C and N by the oxalate reagent, and examined phosphate-uptake kinetics and C:P ratios in adsorbed and internal pools of *Trichodesmium* cultures grown under P-limited or P-replete conditions. The removal efficiency of surface-adsorbed $^{33}PO_4$ (labelled with a single 5-min pulse) from the surface of *Trichodesmium* cultures by the oxalate reagent was ~90% (Supplementary Fig. 1). Our laboratory results also indicated that none of the C and N in phytoplankton was surface-adsorbed, because their levels in the total and intracellular pools

were essentially the same (Supplementary Fig. 2). Our previous results have shown that application of the oxalate reagent does not produce any cell breakage or lysis, or affect the physiological status of the phytoplankton in any way⁸.

The molar C:N:surface-adsorbed P ratios (median and mean) (80–105 C:15–18 N:1 P) of the field-collected *Trichodesmium* colonies were not significantly different from the elemental ratios calculated using the total (= surface-adsorbed + intracellular) P pool (71–80 C:13–14 N:1 P; Fig. 1 and Supplementary Table 1). These results suggest that the Redfield ratio based on total P measurements in plankton largely reflects the contribution of the surface-adsorbed P pool. In contrast, internal nutrient ratios (316–434 C:59–83 N:1 intracellular P) in the same colonies were four to five times higher than in the surface-adsorbed and total fractions, and do not conform with the Redfield model (Fig. 1).

Our study focused on P partitioning in the biogeochemically critical diazotroph *Trichodesmium*, because Redfield hypothesized that the constant elemental ratios in the ocean are a reflection of the chemical composition of N-fixing organisms¹. Our data suggest that that is indeed the case in the surface-adsorbed pool. Our results also showed that surface scavenging of P is not unique to *Trichodesmium*. High levels of surface-adsorbed P were also measured in *Pseudo-*

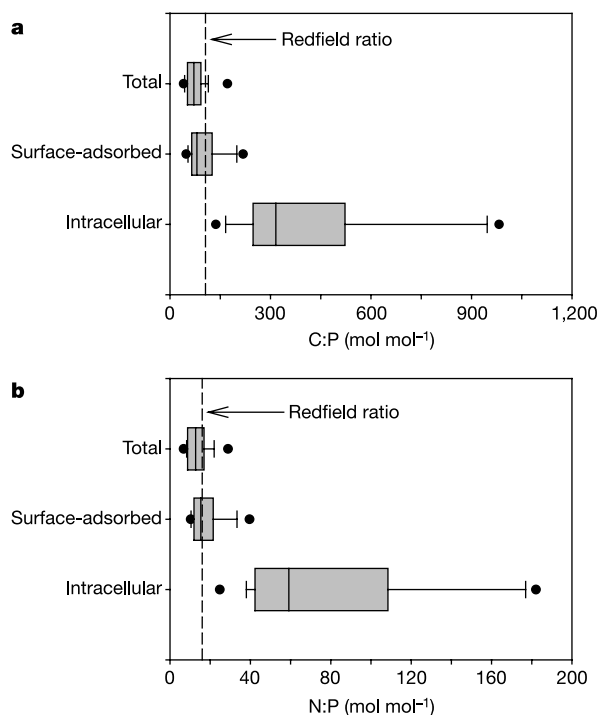


Figure 1 Box plots of elemental composition (C, N, P) in the total, surface-adsorbed and intracellular pools in field-collected *Trichodesmium* colonies from the western subtropical Atlantic Ocean. **a**, C:P ratios. **b**, N:P ratios. The line within the box is the median, and the boundary of the boxes indicates the 25th and 75th percentiles. Error bars to the left and right of the box indicate the 10th and 90th percentiles. Filled circles show outlying points. A *t*-test showed that elemental compositions in the surface-adsorbed and total pools were not significantly different (C:P; *t* = 1.64, *P* > 0.05, N:P; *t* = 1.34, *P* > 0.05).

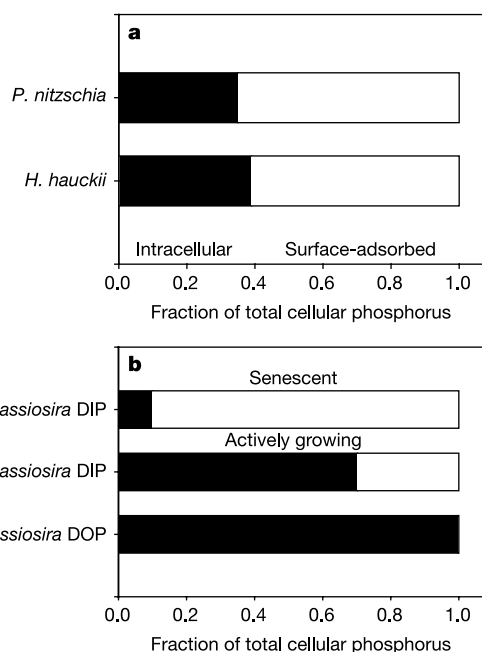


Figure 2 Cellular partitioning of phosphorus in phytoplankton. **a**, Fraction of intracellular and surface-adsorbed P measured in field-collected samples from blooms of *P. nitzschia* sp. and *H. hauckii*. Samples were collected on 23 April 2003 at 06° 53.06' N, 52° 01.95' W and on 26 April 2003 at 09° 48.55' N, 54° 04.95' W. As observed for *Trichodesmium*, more than 60% of the total P measured in those blooms was in the surface-adsorbed pool. **b**, Fraction of intracellular and surface-adsorbed P measured in the centric diatom *Thalassiosira weissflogii* under different growth (senescent versus actively growing) and substrate (DIP versus DOP) conditions in the laboratory. DIP, dissolved inorganic phosphate; DOP, dissolved organic phosphate (in this case glycerophosphate). The amount of surface-adsorbed P (90%) in the senescent diatoms was considerably higher than the amount measured in the exponential growth phase (30%). These results suggest that a fraction of the surface-adsorbed pool is being internalized and used for growth. In contrast to the substantial removal of surface-adsorbed P observed in the culture grown on DIP (senescent and actively growing), none of the cellular P in cultures grown on DOP was removed by the reagent. This is consistent with the tendency of oxyanions such as phosphate to form, by ligand exchange, surface complexes with hydrous oxides in biogenic surfaces and minerals. This process allows removal of surface-adsorbed P by the oxalate wash because it is effective at removing hydrous oxides.

nitzschia sp. and *Hemiaulus hauckii* blooms from two locations in the western Atlantic Ocean, as well as in the diatom *Thalassiosira weissflogii* (Fig. 2), and four other algal taxa in the laboratory (Supplementary Fig. 2).

These results indicate that P, like scavenged cations such as iron^{8,9}, is present in phytoplankton in both surface-adsorbed and intracellular pools. The ratios of these pools in phytoplankton are variable, and do not always reflect Redfield stoichiometry. Although previous studies have also reported large deviations in Redfield ratios^{10–12}, those studies have not addressed the impact of different cellular pools of P.

In P-replete and P-limited laboratory *Trichodesmium* cultures, C:P ratios showed the same trends (Fig. 3a), although the magnitude of differences between pools was somewhat lower than in field samples. The total cellular ratios of P-replete cultures were at near-Redfield values of 95, but intracellular ratios of these same cultures were 1.7 times higher, at 161. The P-limited cultures had higher total

cellular C:P ratios of 207, but intracellular ratios were about 1.4 times greater at nearly 300—very close to the increased intracellular values in the field collections (Fig. 1). These high intracellular C:P ratios in both natural and cultured *Trichodesmium* were consistent with calculated cellular C:P values (range 190–390 mol mol⁻¹), based on RNA as the major intracellular P compound under P-limited conditions^{2,13}.

The existence of two distinct cellular P pools could have significant implications for studies using the Redfield model to define phytoplankton nutrient limitation. For example, whereas total and surface-adsorbed elemental ratios in the *Trichodesmium* colonies suggest P sufficiency (Redfield ratio; N:P range = 10–20), the intracellular pool was depleted in P (higher than Redfield ratio: N:P range = 50–300; Fig. 4a). The current Redfield model based on total P levels may thus not adequately reflect the physiological status of some phytoplankton species. Consistent with that hypothesis, N fixation by *Trichodesmium* in the western subtropical Atlantic Ocean during our study was suppressed at two-thirds of our stations, despite the fact that total elemental ratios were close to the Redfield model (Figs 1 and 4a). Nutrient ratios in the intracellular pool suggested that N fixation was low because those *Trichodesmium* colonies were P-deficient (Fig. 4a), as previously reported for the same region of the Atlantic Ocean^{14,15}.

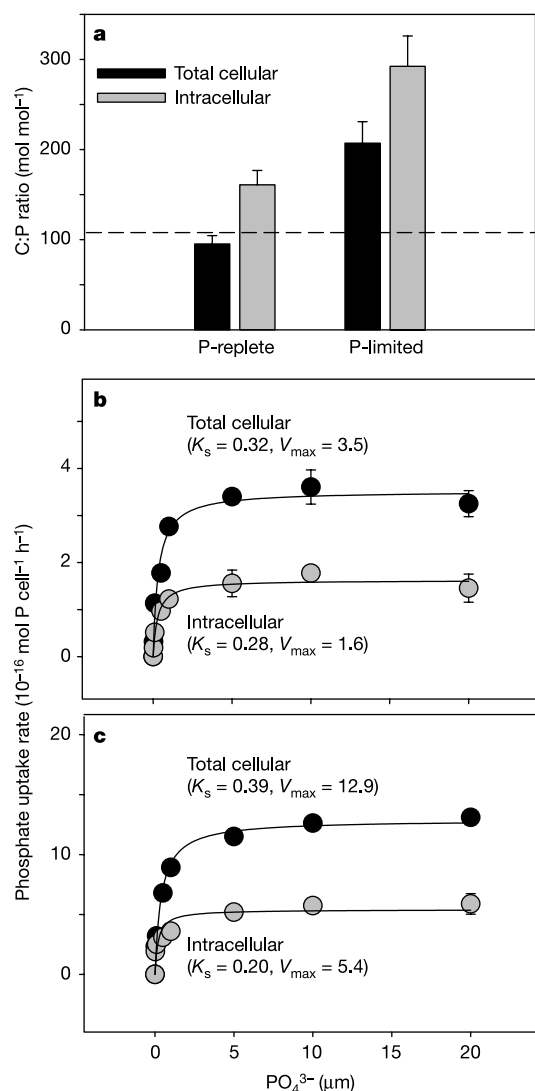


Figure 3 Effects of P-limited and P-replete growth on intracellular (oxalate-washed) and total cellular (seawater-washed) P content and P uptake rates in laboratory cultures of *Trichodesmium* IMS101. Values are the means and error bars are the standard deviations of triplicate cultures. **a**, Molar C:P ratios in both cellular pools in P-limited and P-replete cultures. The dashed line represents the Redfield ratio. **b**, Phosphate uptake rates into total cellular and intracellular pools as a function of phosphate concentration, in the same P-replete cultures shown in **a**. Shown are half-saturation constants (K_s , in μM) and maximum uptake rates (V_{max} , in mol P cell⁻¹ h⁻¹) for both cellular pools. **c**, As in **b**, but for P-limited cultures.

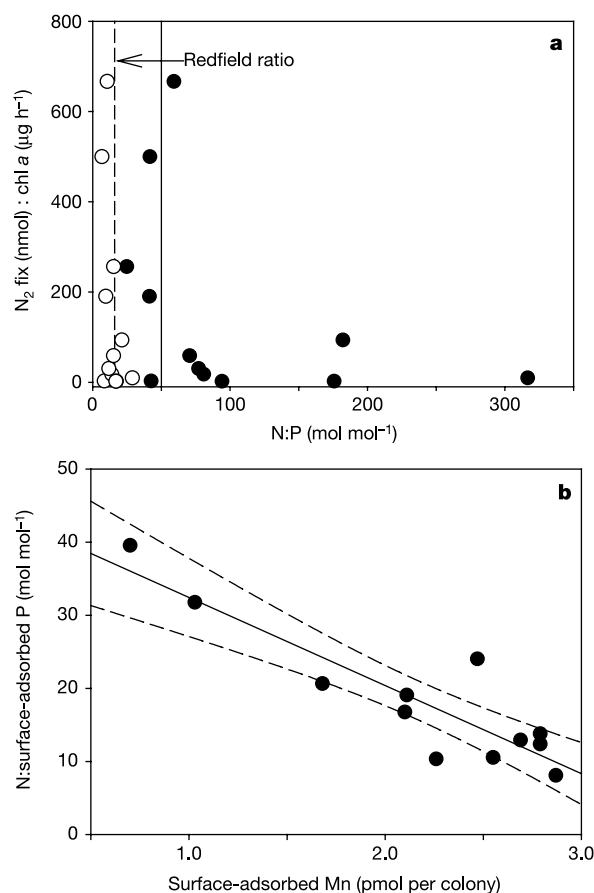


Figure 4 Relationship between different phosphorus pools and nitrogen fixation (**a**) and surface-adsorbed manganese (**b**). **a**, Relationship of N₂ fixation (reported as nmol of C₂H₂ per μg of chlorophyll *a* per hour) with total (open circles) and intracellular (filled circles) N:P ratios measured in the field-collected *Trichodesmium* colonies. Solid line at N:P = 50 is the critical ratio that marks the transition of P-limitation in phytoplankton². **b**, N:surface-adsorbed P ratios versus surface-adsorbed Mn measured in field-collected *Trichodesmium* colonies. Regression equation (solid line): N:surface-adsorbed P = -12.1 × [surface-adsorbed Mn] + 44.5; $r^2 = 0.81$. Dashed lines indicate 95% confidence limits.

Whereas the particle reactivity of P in the environment has been studied by geochemists¹⁶, no previous study has addressed the impact of that mechanism on Redfield stoichiometry and phytoplankton physiology. The existence of the surface-adsorbed and intracellular P pools in phytoplankton suggests that P uptake could be a two-step kinetic process; that is, adsorption to the cell surface, followed by internalization, similar to that reported for iron¹⁷. The different amounts of surface-adsorbed P in senescent (~90%) and in exponentially growing (30%) cultures suggest that phytoplankton have developed mechanisms to access at least some of the surface-adsorbed pool (Fig. 2). However, further studies are needed to evaluate whether surface-adsorbed P can be internalized fast enough to make a significant difference in the nutritional status of different types of phytoplankton.

Our results suggest that kinetic constants for phytoplankton P uptake need to be re-evaluated using intracellular levels of P instead of total concentrations. Removal of the surface-adsorbed P pool with the oxalate reagent results in substantially lower half-saturation constants for uptake (K_s)—88% and 51% of those measured in total cellular pools for P-replete and P-limited cultures, respectively (Fig. 3b, c). Maximum uptake rates (V_{max}) are even more strongly affected, and oxalate-washed values are only 51% (P-replete) and 42% (P-limited) as high as in total cellular pools. Previous physiological P-uptake and growth assessments and N:P requirements of phytoplankton that considered only the total cellular pool^{18,19} may therefore need re-interpretation in relation to intracellular P pools.

Surface scavenging of P by *Trichodesmium* colonies appears to be mediated by the amount of surface-adsorbed manganese (Supplementary Fig. 3). This is consistent with the covariance between total P and Mn concentrations previously reported in biogenic particles from the North Atlantic²⁰. The surface scavenging of phosphate is explained by the tendency of oxyanionic elements to form, by ligand exchange, surface complexes with hydrous oxides on biogenic surfaces and minerals²¹. This mechanism of P adsorption in *Trichodesmium* colonies is further substantiated by the positive relationship between surface-adsorbed Mn and other oxyanions such as molybdate (Supplementary Fig. 3). This adsorption trend suggests that the surface-adsorbed enrichment of phosphate is caused by inorganic scavenging processes, rather than by active biogenic uptake. This is consistent with our laboratory results showing that the oxalate reagent did not remove any of the cellular P from cultures grown on a dissolved organic P source, in contrast to phosphate-grown cultures (Fig. 2b).

Our results showed that the amount of surface-adsorbed manganese on the *Trichodesmium* colonies seems to be the main factor determining the adsorption capacity for P (Supplementary Fig. 3). Therefore, N:P ratios in the colonies were inversely related to levels of surface-adsorbed manganese (Fig. 4b). This relationship suggests that deviations in Redfield stoichiometry (N:P range = 10–40; Fig. 4b) observed in the Atlantic Ocean could be partially explained by the adsorption of P onto cell-surface-bound metal hydroxides and oxides, rather than being attributed solely to N fixation²² or N pollution²³.

Significant decoupling of elemental ratios between the surface-adsorbed and intracellular pools could have far-reaching implications for marine biogeochemistry. For example, current Redfield ratios normalized to total cellular P may not adequately reflect the nutritional status of phytoplankton, as they appear to be largely determined by the adsorption of this nutrient onto reactive chemical surface sites of plankton. Furthermore, models that use Redfield stoichiometry to couple changes in nutrient inventories and marine biogeochemical processes^{24,25}, should take into consideration the existence of the two different P pools in phytoplankton. □

Methods

Trichodesmium colonies were collected using a Zodiac inflatable deployed from the RV

Seward Johnson from 18 April to 20 May 2003 in the subtropical North Atlantic Ocean (Supplementary Table 1). Colonies were collected at a depth of ~5 m by towing an acid-washed 102-µm plankton net. In a class-100 bench, ~200 colonies per station were hand-picked from the acid-washed cod-end using a plastic inoculating loop, deposited into 3-ml acid-washed Teflon vials and digested in a combination of quartz-distilled HCl, HNO₃ and HF. Fe, Mn and Mo content in the *Trichodesmium* colonies were then measured by inductively coupled plasma mass spectrometry (ICP-MS). P concentrations in the digested colonies were determined by spectrophotometric techniques developed for small volumes published elsewhere¹⁴. Particulate organic C and N levels in the colonies were determined using a Carlo Erba NA1500 NCS system. Intracellular levels of P, Fe, Mn and Mo were determined as described above after washing ~100 colonies with the oxalate reagent⁸. The redox potential of oxalic acid ($E_H^0 = -631$ mV) indicates that this reagent reduces most of the iron ($E_H^0 = -295$ mV) and manganese ($E_H^0 = -50$ mV) hydroxides bound to phytoplankton. The surface-adsorbed pool is calculated by the difference between the total and intracellular determinations. Nitrogen fixation was determined on isolated colonies using the protocols described in ref. 14.

Removal efficiency of adsorbed phosphate by the oxalate reagent was tested on laboratory cultures of *Trichodesmium* IMS101 and the centric diatom *Thalassiosira weissflogii* following 5-min-pulse labelling of extracellular pools using $^{33}\text{PO}_4^{3-}$ (Supplementary Fig. 1), as previously described for iron (refs 8, 9). For uptake kinetics determinations, cultures of the Atlantic isolate *Trichodesmium* IMS101 were grown in P-replete (20 µM) or P-limited (0.2 µM) seawater medium, and harvested in early exponential phase to generate short-term (1 h) Michaelis–Menten PO_4^{3-} uptake curves using the tracer $^{33}\text{PO}_4^{3-}$ (ref. 18). Elemental ratios in the cultures were determined using the same protocols described above for the field-collected *Trichodesmium* colonies.

Received 3 September; accepted 21 October 2004; doi:10.1038/nature03125.

- Redfield, A. C. The biological control of chemical factors in the environment. *Am. Sci.* **46**, 205–221 (1958).
- Geider, R. J. & La Roche, J. Redfield revisited: variability of C:N:P in marine microalgae and its biogeochemical basis. *Eur. J. Phycol.* **37**, 1–17 (2002).
- Van Cappellen, P. & Ingall, E. D. Redox stabilization of the atmosphere and oceans by phosphorus-limited marine productivity. *Science* **271**, 493–496 (1996).
- Wang, X. J., Matear, R. J. & Trull, T. W. Nutrient utilization ratios in the Polar Frontal Zone in the Australian sector of the Southern Ocean: A model. *Glob. Biogeochem. Cycles* **17**, 1009, doi:10.1029/2002GB001938 (2003).
- Guildford, S. J. & Hecky, R. E. Total nitrogen, total phosphorus, and nutrient limitation in lakes and oceans: Is there a common relationship? *Limnol. Oceanogr.* **45**, 1213–1223 (2000).
- Arrigo, K. R. *et al.* Phytoplankton community structure and drawdown of nutrients and CO₂ in the Southern Ocean. *Science* **238**, 365–367 (1999).
- Quigg, A. *et al.* The evolutionary inheritance of elemental stoichiometry in marine phytoplankton. *Nature* **425**, 291–294 (2003).
- Tovar-Sanchez, A. *et al.* A trace metal clean reagent to remove surface-bound iron from marine phytoplankton. *Mar. Chem.* **82**, 91–99 (2003).
- Hutchins, D. A., Wang, W. X., Schmidt, M. A. & Fisher, N. S. Dual labeling techniques for trace metal biogeochemical investigations in aquatic plankton communities. *Aquat. Microb. Ecol.* **19**, 129–138 (1999).
- Bury, S. J., Boyd, P. W., Preston, T., Savidge, G. & Owens, N. J. P. Size-fractionated primary production and nitrogen uptake during a North Atlantic phytoplankton bloom: implications for carbon export estimates. *Deep-Sea Res.* **1** **48**, 689–720 (2001).
- Pahlow, M. & Riebesell, U. Temporal trends in deep ocean Redfield ratios. *Science* **287**, 831–833 (2000).
- Schneider, B., Schlitzer, R., Fischer, G. & Nothig, E.-M. Depth-dependent elemental composition of particulate organic matter (POM) in the ocean. *Glob. Biogeochem. Cycles* **17**, 1032, doi:10.1029/2002GB001871 (2003).
- Laws, E. A., Redalje, D. G., Karl, D. M. & Chalup, M. S. A theoretical and experimental examination of the predictions of two recent models of phytoplankton growth. *J. Theor. Biol.* **105**, 469–491 (1993).
- Saudo-Wilhelmy, S. A. *et al.* Phosphorus limitation of nitrogen fixation by *Trichodesmium* in the central Atlantic Ocean. *Nature* **411**, 66–68 (2001).
- Mills, M. M., Ridame, C., Davey, M., La Roche, J. & Geider, R. J. Iron and phosphorus co-limit nitrogen fixation in the eastern tropical North Atlantic. *Nature* **429**, 292–294 (2004).
- Bjerrum, C. J. & Canfield, D. E. Ocean productivity before about 1.9 Gyr ago limited by phosphorus adsorption onto iron oxides. *Nature* **417**, 159–161 (2002).
- Hudson, R. J. M. & Morel, M. M. Trace metal transport by marine microorganisms: implications of metal coordination kinetics. *Deep-Sea Res.* **40**, 129–150 (1993).
- McCarthy, J. J. & Carpenter, E. J. *Oscillatoria* (*Trichodesmium*) *thiebautii* (Cyanophyta) in the central North Atlantic Ocean. *J. Phycol.* **15**, 75–82 (1979).
- Klausmeyer, C. A., Litchman, E., Daufresne, T. & Levin, S. A. Optimal nitrogen-to-phosphorus stoichiometry of phytoplankton. *Nature* **429**, 171–174 (2004).
- Kuss, J. & Kremling, K. Spatial variability of particle associated trace elements in near-surface waters of the North Atlantic (30°N/60°W to 60°N/2°W), derived by large-volume sampling. *Mar. Chem.* **68**, 71–86 (1999).
- Sigg, L. & Stumm, W. The interaction of anions and weak acids with the hydrous goethite (α-FeOOH) surface. *Colloids Surf.* **2**, 101–117 (1981).
- Michaels, A. F. *et al.* Inputs, losses and transformations of nitrogen and phosphorus in the pelagic North Atlantic Ocean. *Biogeochemistry* **35**, 181–226 (1996).
- Fanning, K. A. Influence of atmospheric pollution on nutrient limitation in the ocean. *Nature* **339**, 460–463 (1989).
- Lenton, T. M. & Watson, A. J. Redfield revisited. 1. Regulation of nitrate, phosphate and oxygen in the ocean. *Glob. Biogeochem. Cycles* **14**, 225–248 (2000).
- Karl, D. M. *et al.* Ecological nitrogen-to-phosphorus stoichiometry at station ALOHA. *Deep-Sea Res.* **11** **48**, 1529–1566 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was supported by NSF Chemical and Biological Oceanography and by NOAA Ocean Global Carbon Cycle Program. We are grateful to T. Gunderson and Y. Zhang for technical support.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.A.S.W. (ssanudo@notes.cc.sunysb.edu).

Public health vaccination policies for containing an anthrax outbreak

Ron Brookmeyer¹, Elizabeth Johnson¹ & Robert Bollinger²

¹Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, Maryland 21205, USA

²Division of Infectious Diseases, Johns Hopkins University School of Medicine, 720 Rutland Avenue, Baltimore, Maryland 21205, USA

Concern about biological weapons has raised questions about the most effective public health policies to contain an anthrax outbreak^{1–3}. We developed a probability model to predict the impact of different anthrax antibiotic and vaccination policies. An anthrax outbreak can be significantly contained by minimizing the delay until initiation of antibiotic prophylaxis. However, even if mass distribution of antibiotics is completed within six days of the initial exposure, then at most about 70% of cases can be prevented. Post-exposure vaccination will not significantly increase that prevention rate if adherence to antibiotic regimens is similar or higher than that attained in the 2001 US outbreak⁴. However, post-exposure vaccination can be useful either in shortening the duration of a prolonged antibiotic regimen, in the event of an antibiotic-resistant strain, or if antibiotic adherence rates are very low. Here we show that a mass pre-exposure vaccination programme for the general population would require very high population coverage rates to significantly increase prevention rates from that achieved with targeted and rapid post-exposure prophylaxis programmes.

Public health planners are uncertain how an anthrax vaccine should be used as part of a comprehensive strategy for addressing the bioterrorism threats posed by *Bacillus anthracis*^{1–3}. Critical public health questions include whether an anthrax vaccine should be used before exposure to immunize the general population, or used as a post-exposure prophylaxis. The answers depend on many factors including the safety and efficacy of prolonged antibiotic prophylaxis, adherence to antibiotic regimens, the time delay before post-exposure antibiotic prophylaxis is initiated and vaccine characteristics such as efficacy, safety and time to achieve immunity. The US Food and Drug administration licensed an anthrax vaccine, Anthrax Vaccine Absorbed (AVA) in 1970. Because AVA can require six doses and up to 18 months to achieve maximum protection, research has accelerated to develop and manufacture an improved anthrax vaccine. It is hoped that recombinant protective antigen (rPA) anthrax vaccine could provide maximum protection in three doses over a period of not more than 28 days^{2,5–6}.

The Centers for Disease Control and Prevention recommended that those people exposed in the 2001 US anthrax attacks receive 60 days of antibiotic prophylaxis. Subsequently, it was recommended that people consider extending the antibiotic therapy with or without three doses of the anthrax vaccine^{3,7}. However, few opted to take this additional recommendation, and adherence to the initially recommended 60-day antibiotic regimen was less than 50%⁸. Nevertheless, no persons among the more than 10,000

recommended for prophylaxis therapy developed disease and among them only a handful of cases were prevented⁹. This experience has raised questions about which public health strategies most effectively contain an anthrax outbreak.

A 1-kg release of an anthrax preparation could contain more than 10¹⁴ spores, roughly equivalent to five-billion lethal doses^{10–11}. Although the magnitude of such a release seems daunting, the critical determinants for public health involve not just the size of the release but also dispersal factors, which include the size of the spores, the type of spore preparation (wet or dry), method of release, wind conditions, temperatures, building insulation and the population density of the exposed areas (both inside and outside of buildings)^{2,10–12}. Atmospheric models predict that a 100-kg release could result in between tens of thousands to several million deaths, depending on these factors^{2,11}. In the face of this uncertainty, how does one determine rational public health policy?

Quantitative models have been used recently to inform the policy debate about smallpox vaccination strategies^{13–17}. To evaluate strategies for containing an anthrax outbreak, we developed a probability model that accounts for the dynamics of spore clearance and germination, and describes the effects of pre-exposure vaccination, post-exposure vaccination, antibiotic prophylaxis and antibiotic adherence. We present the results in terms of the percentages of cases that can be prevented with various public health policies. We show that these percentages do not vary appreciably by exposure level (dose of inhaled spores), thus allowing public health policy to be guided, even with the great uncertainty surrounding the specifics of bioterrorism threats.

We considered three models for the dose of inhaled spores: a fixed high-exposure level in which all persons are exposed to the infectious dose ID(50), which yields a rate of 5,000 cases per 10,000

Table 1 Per cent of cases of anthrax prevented by post-exposure intervention

Antibiotic response time*	Antibiotic duration (days)	Antibiotic adherence†	Percentage of cases prevented Post-exposure vaccination‡	
			No	Yes
High exposure§				
None	None	None	0	6
Slow	60	Partial	40	44
Slow	60	Full	46	47
Slow	120	Full	47	47
Rapid	60	Partial	58	64
Rapid	60	Full	68	69
Rapid	120	Full	69	69
Variable moderate exposure§				
None	None	None	0	6
Slow	60	Partial	39	43
Slow	60	Full	46	46
Slow	120	Full	46	46
Rapid	60	Partial	57	63
Rapid	60	Full	67	68
Rapid	120	Full	68	68
Low exposure§				
None	None	None	0	9
Slow	60	Partial	48	52
Slow	60	Full	55	55
Slow	120	Full	55	55
Rapid	60	Partial	66	71
Rapid	60	Full	75	76
Rapid	120	Full	76	76

*Efficacy of antibiotics is assumed to be 100% (that is, the probability of disease protection if on antibiotics at the time of spore germination is 1). With slow response, mass antibiotic distribution begins 6 days after initial exposure and the distribution is completed over the following 6 days (total response time = 12 days). With rapid response, mass antibiotic distribution begins 3 days after initial exposure and distribution is completed over the following 3 days (total response time = 6 days).

†With full adherence, 100% of persons complete the recommended duration of antibiotics (60 or 120 days). With partial adherence to a 60-day regimen, 25% of persons each complete 15, 30, 45 and 60 days of antibiotics.

‡Post-exposure vaccine protection (with 95% efficacy) is assumed to begin 28 days after persons begin antibiotics. In the case of no antibiotics, post-exposure vaccine protection begins 28 days after the sixth day following exposure.

§High exposure is the ID(50). Variable moderate exposure assumes that the percentages exposed to the ID(90), ID(50), ID(10) and ID(1) are respectively 2%, 10%, 28% and 60%, with an average about the ID(10) of 1,020 cases out of 10,000 exposed, if no intervention. Low exposure is the ID(1).

exposed persons with no public health intervention; a fixed low-exposure level in which all persons are exposed to the ID(1) and which yields a rate of 100 cases per 10,000 exposed; and a variable moderate exposure level in which the numbers of inhaled spores varied across the population and yields an average rate of 1,020 cases per 10,000 exposed.

Table 1 displays the percentage of cases prevented with various post-exposure interventions. Surprisingly, the results do not vary appreciably by exposure level. For example, with a rapid 6-day response in which mass antibiotic distribution begins 3 days after the initial exposure and is completed after an additional 3 days, the percentage of cases prevented with full 60-day antibiotic adherence were 68%, 67% and 75% for the high, variable moderate and low exposures, respectively. With a slower 12-day response in which mass antibiotic distribution begins 6 days after the initial exposure and is completed after an additional 6 days, the corresponding percentages of cases prevented were 46%, 46% and 55%. Figure 1 shows the relationship between the percentage of cases prevented and the delay until beginning mass antibiotic distribution. The figure shows that prevention rates, which do not vary significantly by exposure level, are high with short delays, but fall to under 50% after a 10-day delay.

We considered the situation when there was only partial adherence to the antibiotic regimen; specifically, when 25% of people completed either 15, 30, 45 or 60 days of a 60-day regimen, similar to that observed in the 2001 US anthrax outbreak⁸. The percentages of cases prevented drops by about 7–10% compared with full antibiotic adherence (Table 1).

We evaluated the preventive value of post-exposure vaccination as an adjuvant to antibiotic prophylaxis. Our first set of calculations assumed a 95% vaccine efficacy with a 28-day delay to achieve vaccine-induced immunity⁵. If people adhere to the antibiotic regimen, then addition of post-exposure vaccination increases the per cent prevented by, at most, 1% (Table 1). If people only partially adhere to the antibiotic regimen, post-exposure vaccination modestly increases the prevention rate by no more than 6% (Table 1). If the strain of *B. anthracis* is resistant to antibiotics, then post-exposure vaccination prevents between 6% and 9% of cases depending on the exposure level (row 1, Table 1). These modest preventive effects of post-exposure vaccination are explained by the time delay to achieve vaccine-induced immunity.

We performed a sensitivity analysis to the time delay to achieve vaccine-induced immunity, to identify characteristics that would improve a vaccine's utility as a post-exposure prophylaxis. Table 2

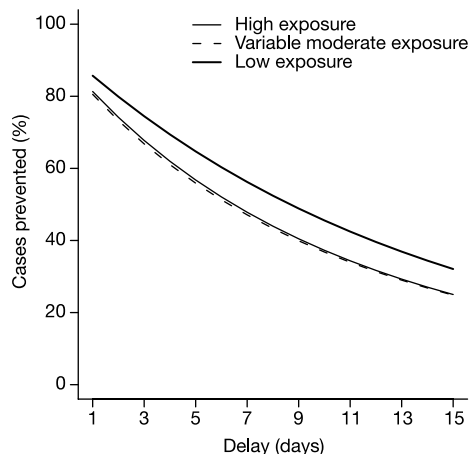


Figure 1 Percentage of cases prevented versus delay of initiating mass distribution of antibiotics. Calculations based on completing mass distribution of antibiotics over 3 days, and assuming that persons adhere completely to a 60-day regimen and that antibiotic efficacy is 100%. Exposure levels are defined in Table 1.

Table 2 Sensitivity of prevention rates to post-exposure vaccine immunogenicity

Cumulative vaccine protection* (%)				Percentage of cases prevented Antibiotic response†				
Day				No adherence 0% efficacy	Speed: slow		Speed: rapid	
7	14	21	28		Partial adhere 90% efficacy	Partial adhere 100% efficacy	Partial adhere 100% efficacy	
High exposure‡								
0	0	0	0	0	36	40	58	
0	0	0	95	6	40	44	64	
0	0	25	95	7	41	44	65	
0	25	50	95	10	42	45	66	
25	50	75	95	17	43	46	67	
Low exposure‡								
0	0	0	0	0	43	48	66	
0	0	0	95	9	48	52	71	
0	0	25	95	10	48	53	72	
0	25	50	95	14	49	53	73	
25	50	75	95	22	51	54	74	

* Cumulative per cent of persons fully protected by vaccine following days 7, 14, 21 and 28 after receiving the first dose of vaccine. As described in footnotes of Table 1, first vaccine dose is given when antibiotics are initiated, or in the case of no antibiotics on day 6 following exposure.

† Antibiotic efficacy (% of persons in whom antibiotics prevented disease) is 0% (as in case of antibiotic-resistant strain), 90% or 100%. Speed in distributing antibiotics and adherence rates are described in footnotes of Table 1.

‡ See footnotes to Table 1.

shows the percentages of cases prevented under different assumptions of vaccine immunogenicity. For example, we considered a highly immunogenic vaccine in which 25%, 50%, 75% and 95% of people achieve immunity within 7, 14, 21 and 28 days following the first vaccine dose, respectively (last row, Table 2). We find that the addition of such a vaccine to an effective antibiotic regimen results in an increase in the prevention rate of no more than 9% in all situations considered (including one in which antibiotics were only partially (90%) effective). A vaccine that also reduced disease severity could have an additional benefit of increasing the percentages of deaths prevented. For example, suppose people are exposed to high doses of an antibiotic-resistant strain; if a single dose of a highly immunogenic vaccine given before disease onset also reduced the mortality rate among cases by 50%, 25%, 10% and 0%, then the percentages of deaths prevented are 25%, 21%, 18% and 17%, respectively.

Prolonged antibiotic prophylaxis may be indicated if either exposure levels to spores are high, or there is significant risk of re-exposure from either re-aerosolization of spores during remediation of contaminated areas or from another bioterrorist event. Shortening a long course of antibiotics may be critical because of difficulty in maintaining adherence, limited supplies of antibiotics, resistance and adverse events associated with long-term

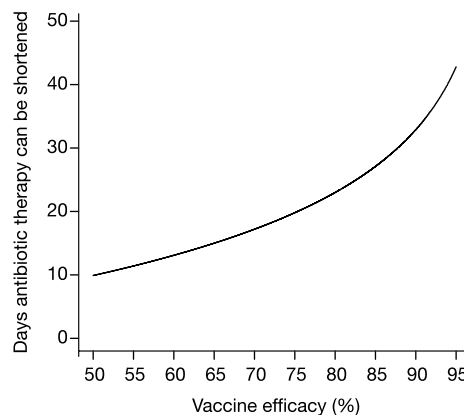


Figure 2 Days that antibiotic therapy can be shortened by post-exposure vaccination versus vaccine efficacy. Assumes the probability that persons are protected from disease by vaccine is the vaccine efficacy and that protection begins at or before cessation of antibiotics. Antibiotic efficacy is assumed to be 100%.

antibiotic therapy. We examined the trade-offs associated with shortening the antibiotic regimen by replacing antibiotic protection with post-exposure vaccination. Vaccination confers long-term immunity, whereas antibiotics are effective for only as long as they are being taken. However, that advantage of vaccination must be balanced by the fact that vaccine efficacy is not 100%. We calculated how many days an antibiotic regimen could be shortened by post-exposure vaccination while not increasing total disease risk, assuming antibiotic efficacy was 100%. The results depend on the vaccine efficacy (Fig. 2). For example, if the vaccine efficacy is 95%, 90% and 80%, then antibiotic regimens could be shortened 43, 33 and 23 days, respectively, provided that the level of vaccine efficacy is achieved before antibiotics are terminated. For example, a 120-day antibiotic regimen could be shortened to only 77 days by using a vaccine that was 95% efficacious.

Table 1 suggests that post-exposure antibiotics can prevent at most about 70% of cases. We calculated whether a mass pre-exposure vaccination programme of the general population could significantly increase that prevention rate. Table 3 shows levels of pre-exposure vaccine coverage in the general population that would be necessary to increase the percentages of cases prevented to either 75% or 90%. The results show that the required level of vaccine coverage depends on multiple factors including exposure level, antibiotic response time and adherence; nevertheless, some general conclusions are possible. To prevent 90% of cases, at the very least 63% of the population would need to receive pre-exposure vaccination. That minimal level of vaccine coverage applies when the exposure is low with rapid response and full antibiotic adherence. Generally, vaccine coverage levels that are considerably higher than 63% are required when the response is slower and antibiotic adherence is lower (Table 3).

Table 3 Pre-exposure vaccination to supplement post-exposure intervention

Antibiotic policy		Vaccination policy		
Response time	Adherence	Post-exposure	Pre-exposure coverage* (%)	
			To prevent 75%	To prevent 90%
High exposure				
None	None	No	79	95
None	None	Yes	77	94
Slow	Partial	No	61	88
Slow	Partial	Yes	56	85
Slow	Full	No	56	86
Slow	Full	Yes	56	85
Rapid	Partial	No	43	81
Rapid	Partial	Yes	31	76
Rapid	Full	No	24	73
Rapid	Full	Yes	20	71
Variable moderate exposure				
None	None	No	79	95
None	None	Yes	77	94
Slow	Partial	No	62	88
Slow	Partial	Yes	59	87
Slow	Full	No	57	86
Slow	Full	Yes	56	86
Rapid	Partial	No	44	81
Rapid	Partial	Yes	33	76
Rapid	Full	No	26	74
Rapid	Full	Yes	23	72
Low exposure				
None	None	No	79	95
None	None	Yes	76	94
Slow	Partial	No	55	85
Slow	Partial	Yes	50	83
Slow	Full	No	47	82
Slow	Full	Yes	47	82
Rapid	Partial	No	28	74
Rapid	Partial	Yes	14	69
Rapid	Full	No	4	65
Rapid	Full	Yes	0	63

Calculations use parameter values given in Table 1 footnotes. Antibiotic regimen is 60 days. Vaccine efficacy equals 95%.

*Percentage of population requiring pre-exposure vaccination necessary to prevent 75% and 90% of cases.

Our analysis shows that minimizing the delay until initiation of antibiotic prophylaxis is key to containing an anthrax outbreak, and reinforces the results of Wein *et al.*¹¹. Heightened clinical awareness of symptoms, improved public health surveillance, and efficient systems for mass antibiotic distribution^{11,18–19} are all critical components in reducing response time. In the 2001 US outbreak, postal workers did not receive antibiotics until more than 9 days after exposure. However, even if mass antibiotic distribution is completed within 6 days of exposure, at most about 70% of cases could be prevented. Our analysis shows that post-exposure vaccination should not be expected to significantly increase that percentage because of the delays to achieve vaccine protection, but nevertheless can be useful in shortening a prolonged antibiotic regimen. Although mass pre-exposure vaccination programmes of the general population do increase prevention rates, very high vaccine coverage rates are required.

These analyses assumed that there are adequate supplies of antibiotics and vaccine, and that all exposed persons needing post-exposure prophylaxis are completely identified. However, if only a fraction, γ , of exposed persons are identified, then the prevention rates in Table 1 would be overestimated and should then be multiplied by γ . In the event of a large bioterrorism attack, γ could well be less than 1. The analyses assumed that implementation of post-exposure response policies, including identification of exposed persons, does not depend on the numbers of exposed persons. Wein and colleagues¹¹ considered models that account for the attack size and the queues that would result from limited resources for an emergency response. Developing bioterrorism response policies is challenging because of limited available data on anthrax outbreaks. Although we cannot plan for all possible events of bioterrorism, having public health policy guidelines in place for rapid and rational decision-making are critical elements of public health preparedness. □

Methods

A probability model based on a competing risks formulation of spore clearance and germination²⁰ was developed to describe the impact of disease control strategies. The risk that a spore germinates is called λ , and the risk that a spore is naturally cleared from the lung by natural mechanisms—including being expelled through a bronchus, swallowed or killed by macrophages—is called the clearance rate θ . These risks are the hazard rates, expressed in units of per day. In addition to these natural clearance mechanisms, a person can be protected from disease by antibiotics or vaccine. If antibiotics are circulating in the body at the time a spore begins to germinate, or if the person has been vaccinated, then it is assumed that the germinating spore is destroyed with probabilities determined by the efficacies of the antibiotics or vaccine. If a spore successfully evades the natural clearance mechanisms, germinates and is not destroyed by vaccine-induced immunity or antibiotics, then it multiplies rapidly, and produces toxin and disease^{21–25}.

Impact of post-exposure prophylaxis

Suppose protection from disease begins at time t_1 then stops at t_2 , and perhaps begins again at t_3 , where the times are measured in days from exposure. Intermittent protection from disease could result from starting and stopping antibiotics, and achieving vaccine-induced immunity.

The probability R_1 that a spore germinates in the interval $(0, t_1)$ is given by:

$$R_1 = \int_0^{t_1} \lambda e^{-(\lambda+\theta)u} du = \frac{\lambda}{\lambda+\theta} [1 - e^{-(\lambda+\theta)t_1}]$$

The probability R_2 that a spore germinates in the interval (t_1, t_2) is equal to 0. The probability R_3 that a spore germinates in the interval (t_2, t_3) is given by:

$$R_3 = \int_{t_2}^{t_3} \lambda e^{-(\lambda+\theta)u} du = \frac{\lambda}{\lambda+\theta} [e^{-(\lambda+\theta)t_2} - e^{-(\lambda+\theta)t_3}]$$

The probability R_4 that a spore germinates after t_3 is equal to 0.

Suppose a person inhales D spores, and let X be the number of spores that will ever germinate in that individual. Assuming X has a Poisson distribution with mean $\mu = D(R_1 + R_2 + R_3 + R_4)$, then the cumulative probability that at least one spore germinates is $c = 1 - e^{-\mu}$, which is:

$$c = 1 - \exp \left\{ -D \frac{\lambda}{\lambda+\theta} [1 - e^{-(\lambda+\theta)t_1} + e^{-(\lambda+\theta)t_2} - e^{-(\lambda+\theta)t_3}] \right\}$$

The estimates of θ from animal data are approximately 0.07 per day, that is, spores are cleared from the lung at about 7% per day^{20,26}. That estimate of the clearance rate is consistent with human data on the incubation period from the largest documented human anthrax outbreak in Sverdlovsk^{10,20} and observations from the 2001 US outbreak²⁰. Estimates of λ , based on estimates of the ID(50) in experimental studies with rhesus

macaque monkeys, suggest that λ ranges between 3.2×10^{-6} and 8.1×10^{-5} per day^{27,28}. Thus, λ is much smaller than θ . The dose to cause disease with probability $p = P/100$ is called the infectious dose ID(P) and is given by²⁰ $D = -(\lambda + \theta)\ln(1 - p)/\lambda$. Substituting this expression for D into the formula for c and approximating $\lambda + \theta \approx \theta$, we find

$$c = 1 - (1 - p)^{(1 - e^{-\theta t_1} + e^{-\theta t_2} - e^{-\theta t_3})} \quad (1)$$

Thus, we have the surprising mathematical result that it is not necessary to have a precise value for λ to obtain our key results.

We divide the population into subgroups in which the periods of disease protection are the same for all persons in a subgroup. For example, consider a subgroup in which antibiotic protection begins at time t_1 and stops at time t_2 , then we set t_3 to infinity in equation (1). Suppose further that post-exposure vaccine protection is achieved at t_v , which occurs before antibiotic protection stops ($t_v < t_2$), then t_2 and t_3 are set to infinity. If vaccine protection occurs after antibiotic protection stops then t_3 is set equal to t_v . The cumulative probability of disease in the population (c) is then a weighted average of the values of c for each subgroup, where the weights are the proportions of people in subgroups. These proportions are determined by the distributions of times in which vaccine immunity is achieved, and the antibiotic start and stop dates (determined by antibiotic adherence rates and efficacy). Among persons exposed to the ID(P), the percentage of cases prevented with post-exposure interventions is:

$$\left(1 - \frac{c}{p}\right)100 \quad (2)$$

Impact of pre-exposure vaccination

The impact of adding a pre-exposure vaccination programme is to reduce the cumulative probability of disease from c to $c(1 - \beta\phi)$, in which ϕ is the vaccine efficacy (that is, the probability that the vaccine protects from disease), and β is the vaccine coverage (that is, the fraction of the population that receives pre-exposure vaccination). The fraction r of all cases prevented by adding a pre-exposure vaccination programme to the post-exposure prophylaxis strategy among persons exposed to the ID(P), is:

$$r = 1 - \frac{c(1 - \beta\phi)}{p} \quad (3)$$

If we solve equation (3) for β we obtain $\beta = \{1 - [(p(1 - r))/c]\phi\}^{-1}$.

The above equation for β was used to calculate the vaccine coverage rates given in Table 3 with $r = 0.75$ and 0.90 .

Shortening antibiotic regimen by post-exposure vaccine

Consider two situations. In situation 1, antibiotics stop at time t_v , at which time vaccine protection begins with probability ϕ . In situation 2, antibiotics are continued for a longer period of time, until t_2 , but there is no vaccine protection. We calculate by how much time antibiotics can be shortened ($t_2 - t_v$) in order that the disease risks in the two situations are equal. We set the two probabilities of disease conditional on being disease free at t_v equal to each other. Solving that equation and approximating $\lambda + \theta \approx \theta$, we obtain $t_2 - t_v = [-\ln(1 - \phi)]/\theta$; this was used to produce Fig. 2.

Received 14 September; accepted 4 October 2004; doi:10.1038/nature03087.

1. Inglesby, T. V. *et al.* Anthrax as a biological weapon: medical and public health management. *J. Am. Med. Assoc.* **281**, 1735–1745 (1999).
2. Inglesby, T. V. *et al.* Anthrax as a biological weapon. Updated recommendations for management. *J. Am. Med. Assoc.* **287**, 2236–2252 (2002).
3. Centers for Disease Control. Notice to readers: use of anthrax vaccine in response to terrorism: supplemental recommendations of the Advisory Committee on Immunization Practices. *Morb. Mortal. Wkly Rep.* **51**, 1024–1026 (2002).
4. Jernigan, J. A. *et al.* Bioterrorism-related inhalation anthrax: The first 10 cases reported in the United States. *Emerg. Infect. Dis.* **17**, 933–944 (2011).
5. Joellenbeck, L., Zwanziger, L., Durch, J. & Strom, B. (eds) *The Anthrax Vaccine: Is it Safe? Does it Work?* (National Academy Press, Washington DC, 2002).
6. National Institute of Allergy and Infectious Diseases. National Institute of Allergy and Infectious Diseases biodefense research agenda for CDC category A agents. (<http://www.niaid.nih.gov/publications/bioterrorism.htm>) (2002).
7. Centers for Disease Control. Update: investigation of bioterrorism-related anthrax. *Morb. Mortal. Wkly Rep.* **50**, 1008–1010 (2001).
8. Shepard, C. W. *et al.* Antimicrobial post-exposure prophylaxis for anthrax: Adverse events and adherence. *Emerg. Infect. Dis.* **8**, 1124–1132 (2002).
9. Brookmeyer, R. & Blades, N. Prevention of inhalational anthrax in the U.S. outbreak. *Science* **295**, 1861 (2002).
10. Meselson, M. *et al.* The Sverdlovsk anthrax outbreak of 1979. *Science* **266**, 1202–1208 (1994).
11. Wein, L. M., Craft, D. L. & Kaplan, E. H. Emergency response to an anthrax outbreak. *Proc. Natl Acad. Sci. USA* **100**, 4346–4351 (2003).
12. Committee on Research and Development Needs for Improving Civilian Medical Response to Chemical and Biological Terrorism Incidents. *Chemical and Biological Terrorism* (National Academy Press, Washington DC, 1999).
13. Kaplan, E. H., Craft, D. L. & Wein, L. M. Emergency response to a smallpox attack: the case for mass vaccination. *Proc. Natl Acad. Sci. USA* **99**, 10935–10940 (2002).
14. Bozette, S. *et al.* A model for a smallpox-vaccination policy. *N. Engl. J. Med.* **348**, 416–425 (2003).
15. Halloran, M. E., Longini, I. M. Jr, Nizam, A. & Yang, Y. Containing bioterrorist smallpox. *Science* **298**, 1428–1432 (2002).
16. Ferguson, N. M. *et al.* Planning for smallpox outbreak. *Nature* **425**, 681–685 (2003).
17. Fraser, C., Riley, S., Anderson, R. M. & Ferguson, N. M. Factors that make an infectious disease outbreak controllable. *Proc. Natl Acad. Sci. USA* **101**, 6146–6151 (2004).
18. Kaufmann, A., Meltzer, M. & Schmid, G. The economic impact of a bioterrorist attack: Are prevention and post attack intervention programs justifiable? *Emerg. Infect. Dis.* **3**, 83–94 (1997).

19. Centers for Disease Control and Prevention. Mass antibiotic dispensing: A primer. (<http://www.phppo.cdc.gov/phtn/antibiotic/default.asp>) (June 2004).
20. Brookmeyer, R., Johnson, E. & Bollinger, R. Modeling the optimum duration of antibiotic prophylaxis in an anthrax outbreak. *Proc. Natl Acad. Sci. USA* **100**, 10129–10132 (2003).
21. Brachman, P. S. Inhalation anthrax. *Ann. NY Acad. Sci.* **353**, 83–93 (1980).
22. Ross, J. M. The pathogenesis of anthrax following the administration of spores by the respiratory route. *J. Pathol. Bacteriol.* **73**, 484–494 (1957).
23. Friedlander, A. M. *et al.* Post-exposure prophylaxis against experimental inhalation anthrax. *J. Infect. Dis.* **167**, 1239–1242 (1993).
24. Friedlander, A. M. Anthrax: clinical features, pathogenesis and potential biological warfare threat. *Curr. Clin. Top. Infect. Dis.* **20**, 335–349 (2000).
25. Cieslak, T. J. & Edward, M. E. Clinical and epidemiological principles of anthrax. *Infect. Dis.* **5**, 51–55 (1999).
26. Henderson, D. W., Peakcock, S. & Belton, F. C. Observations on the prophylaxis of experimental pulmonary anthrax in the monkey. *J. Hyg. (Lond.)* **54**, 28–35 (1956).
27. Haas, C. N. On the risk of mortality to primates exposed to anthrax spores. *Risk Anal.* **22**, 189–193 (2002).
28. Watson, A. & Keir, D. Information on which to base assessments of risk from environmental contaminated with anthrax spores. *Epidemiol. Infect.* **113**, 479–490 (1994).

Acknowledgements The authors acknowledge the comments of the Anthrax Modeling Working Group of the Secretary's Council on Public Health Preparedness of the Department of Health and Human Services. This research was partially funded by the Fogarty International Center and a grant from the National Institute of Allergy and Infectious Diseases.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.B. (rbrook@jhsph.edu).

Transmissibility of 1918 pandemic influenza

Christina E. Mills¹, James M. Robins^{1,2} & Marc Lipsitch^{1,3}

¹Department of Epidemiology, ²Department of Biostatistics, and ³Department of Immunology and Infectious Diseases, Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115, USA

The 1918 influenza pandemic killed 20–40 million people worldwide¹, and is seen as a worst-case scenario for pandemic planning. Like other pandemic influenza strains, the 1918 A/H1N1 strain spread extremely rapidly. A measure of transmissibility and of the stringency of control measures required to stop an epidemic is the reproductive number, which is the number of secondary cases produced by each primary case². Here we obtained an estimate of the reproductive number for 1918 influenza by fitting a deterministic SEIR (susceptible-exposed-infectious-recovered) model to pneumonia and influenza death epidemic curves from 45 US cities: the median value is less than three. The estimated proportion of the population with A/H1N1 immunity before September 1918 implies a median basic reproductive number of less than four. These results strongly suggest that the reproductive number for 1918 pandemic influenza is not large relative to many other infectious diseases³. In theory, a similar novel influenza subtype could be controlled. But because influenza is frequently transmitted before a specific diagnosis is possible and there is a dearth of global antiviral and vaccine stores, aggressive transmission reducing measures will probably be required.

The emergence of a new pandemic influenza subtype is expected³. Pandemics have occurred regularly throughout this century (1918, 1957 and 1968)⁴, and opportunities for the generation of new subtypes⁵ have persisted and probably expanded³. Pandemic influenza spreads rapidly, has high attack rates and kills millions of people worldwide^{1,4,6}. The 1918 pandemic was particularly destructive. The case fatality proportion (CFP) was ten times higher than in

all other influenza pandemics and was unusually high in young adults^{4,7}.

Understanding the great speed with which influenza is transmitted is crucial to effective preparation for pandemics. The doubling time for an epidemic curve (about three days in the 1918 pandemic) is a function of the reproductive number (R) and the serial interval (ν), which is the average time between a primary and secondary case. The magnitude of R determines the intensity of measures required to halt transmission. The components of ν , that is, the duration of the latent and infectious periods, determine how and when these measures must be applied. Estimates of R for pandemic influenza vary widely, ranging from 1.68 to 20 (refs 8–12), and are thus of limited value for pandemic preparation. Furthermore, R depends on both the infectious agent and the host population; for example, estimates for measles vary between rural and urban populations². To our knowledge, there are no estimates of the range of R values that might be expected if a strain similar to 1918 pandemic influenza emerges.

We estimated R by fitting a deterministic SEIR model with homogeneous mixing to the excess pneumonia and influenza (P&I) death curves for 45 cities (Fig. 1). We assumed distributions of infectiousness consistent with previous studies⁸ and with viral shedding data⁴, giving mean latent and infectious periods of 1.9 days and 4.1 days, respectively. Infections in the SEIR model were transformed to P&I mortality by assuming a 2% CFP^{7,13}. Time to death was based on influenza autopsy reports^{14–16}, with a mean survival time of two weeks (see Supplementary Information). Excess deaths were defined as the deaths above the median for 1910–16 (described previously¹⁷, see Fig. 1 and Supplementary Information). The choice of baseline does not substantially affect the results; the 1918 influenza CFP was so much higher than for previous epidemics that any baseline is a small proportion of the 1918 P&I death curve (see Supplementary Information).

We estimate that R for 1918 pandemic influenza was approximately 2–3 (Fig. 2). The median estimated R for 45 cities was 2 (interquartile range 1.7, 2.3), based on the first three weeks of each epidemic curve with greater than one excess P&I death per 100,000 population. To address the possibility that these initial R estimates were biased downward—the model ignores the possibility of substantial slowing of exponential growth due to depletion of susceptible hosts, heterogeneous mixing and/or transmission-reducing interventions—we also fit, for each city, the two adjacent weeks of data with the greatest exponential growth rate. The median estimated R (2.7) increased and the interquartile range (2.3, 3.4) widened. The maximum initial (6.3) and extreme (6.5) R estimates were similar. The extreme R values serve as an upper bound; R estimates based on fits of other regions of the epidemic curves tended to be lower than the extreme R values (see Supplementary

Information). These results strongly support the conclusion that the estimated R for 1918 pandemic influenza was not large relative to that for many other infectious diseases².

The initial R estimate was not significantly correlated with factors that included latitude, longitude, population size¹⁸, population density^{19,20}, age distribution²¹ or sex distribution²¹. These results are consistent with contemporary findings by Pearl, who sought correlates of an “epidemic index”²¹. The extreme R estimates were weakly positively correlated with population density in 1910 (Spearman correlation $\rho = 0.5$, $P = 0.001$) and 1920 ($\rho = 0.4$, $P = 0.011$), but only the former association remains statistically significant after the Bonferroni correction for multiple comparisons.

Our R estimates were based on excess P&I mortality data, which are surrogates for influenza case data. We assessed the potential bias of using mortality data in several ways. Sensitivity analyses indicated that the assumptions governing the case-to-death transformation in our model (the CFP and the time-to-death distribution) did not affect the magnitude of R estimates (see Supplementary Information). Furthermore, calculations described in the Supplementary Information indicate that mortality data will quickly (within two weeks of the start of the epidemic) reach exponential growth at a rate determined by R , even if transmission and mortality are concentrated in separate groups. Thus, mortality-based R estimates should not be biased. We also estimated R using symptomatic influenza reports from three of the 45 cities. They are consistent with these theoretical results, and are comparable to mortality-based estimates (see Supplementary Information).

It is unknown how similar the recent and 1918 pandemic A/H1N1 influenza viral shedding patterns are, nor is it clear how these patterns translate into transmission probability. To better understand the dependence of R on assumptions about the serial interval and its components, we plotted R estimates based on a linearized SEIR model with an exponential growth rate corresponding to the median rate found in calculating the extreme values of R (Fig. 3, see Supplementary Information). R is less than 3 for most plausible mean ν , regardless of the fraction of ν that is spent in the latent period. To observe an R above 5, the mean ν for 1918 influenza would have had to be several times longer than that of more recent influenza strains⁴.

The proportion of the population susceptible at the start of the pandemic determines the relationship between R and the basic reproductive number (R_0), which is the number of secondary cases generated by a primary case in a completely susceptible population². Frost hypothesized that a 1918 pandemic-like strain spread throughout America in the spring of 1918 (ref. 22), and recent analyses support this ‘herald wave’ hypothesis²³. Anecdotal evidence suggests that those who fell ill in the spring were protected from

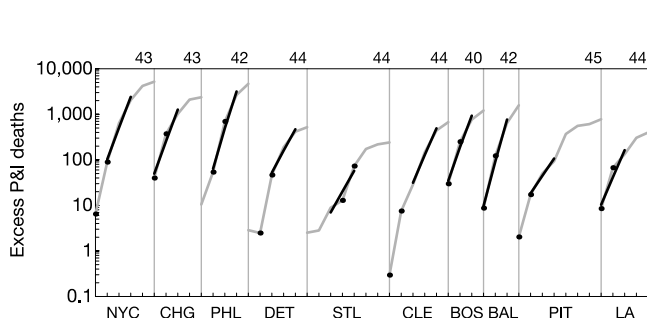


Figure 1 Graph of the logarithm of excess P&I deaths in 1918 for the ten most populous cities in the US. Curves from each city are separated by vertical bars for clarity, with the week number listed above each peak. Curves are shown from left to right, in order of decreasing population size: New York City (NYC), Chicago (CHG), Philadelphia (PHL), Detroit (DET), St Louis (STL), Cleveland (CLE), Boston (BOS), Baltimore (BAL), Pittsburgh (PIT) and Los Angeles (LA). Raw data are shown as grey lines. Black lines indicate model fits for the initial R estimates. Black dots indicate the weeks used for the extreme R estimates.

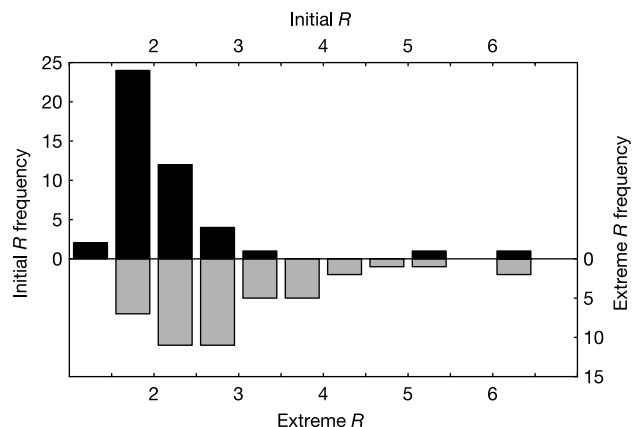


Figure 2 Histogram of initial and extreme estimated R values for 45 cities during the 1918 influenza pandemic. Dark bars show initial R estimates, grey bars show extreme R estimates.

disease in the autumn pandemic²⁴. Nevertheless, a large majority of the population was probably susceptible to the A/H1N1 pandemic strain in September 1918. In a typical epidemic transmission season, 15–25% of the population becomes infected with influenza⁴. The herald wave is believed to have arrived late in the 1917–18 transmission season. Using 70% as a conservative lower bound for the fraction susceptible at the start of the autumn pandemic, the medians for our initial and extreme R_0 are 2.9 and 3.9.

We have presented a range of R estimates for 1918 pandemic influenza in 45 US cities. We know of one previous R estimate for 1918 pandemic influenza, obtained by fitting the entirety of a single curve of influenza deaths from cities across England and Wales¹⁰. Although the previous estimate was described by its author as having a poor fit to data¹⁰, and despite the additional assumptions implicit in that analysis, it is reassuring that our R estimates for 1918 A/H1N1 are similar. The median initial and extreme R values are slightly higher than 1957 A/H2N2 and 1968 A/H3N2 pandemic R estimates^{8,11} (see Supplementary Discussion), but are well below that of many other infections². It is likely that an exceptionally high CFP in young adults, rather than an unusually high R , is what distinguishes the epidemic curve of 1918 from more recent pandemics.

For many infectious agents, explosive epidemics indicate high transmissibility. While the 1918 pandemic progressed quite rapidly, our results indicate that 1918 pandemic influenza A/H1N1 was not highly transmissible relative to other influenza subtypes^{8,11} and infectious agents². A similar pandemic with an R_0 of 2–4 could in principle be prevented by vaccinating or administering antiviral prophylaxis to 50–75% of the population, a figure that may require some adjustment owing to heterogeneous mixing patterns in real populations²⁵. Unfortunately, controlling a future pandemic will not be so simple. At present, vaccine production capacity and antiviral medication stockpiles are insufficient to provide such broad coverage, even in wealthy countries³, although judicious use of available supplies may nevertheless reduce mortality during a pandemic⁸. Control measures based on case identification (for example, contact tracing, isolation, targeted prophylaxis and treatment) that were central to the control of severe acute respiratory syndrome (SARS), whose R was similar to that for 1918 influenza^{26–28}, will only be partially successful for influenza¹². For influenza, unlike SARS, substantial transmission occurs before the onset of case-defining symptoms¹². This implies that measures that generally reduce contacts between persons, regardless of infection status, may be our most powerful protection against a pandemic until adequate vaccine and antiviral medicines can be produced, at which point mass-vaccination and prophylaxis may be more effective than targeted approaches.

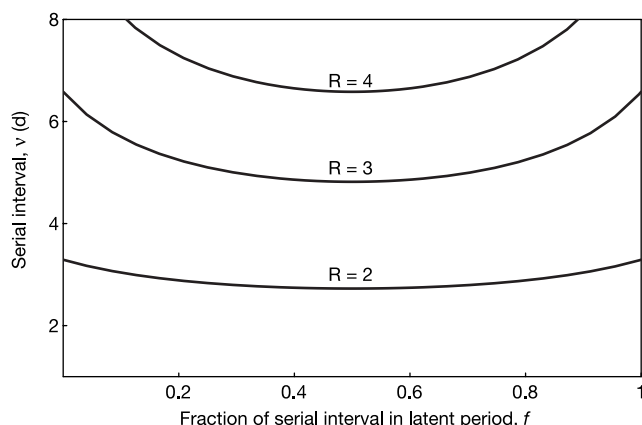


Figure 3 Graph of the relationship between the serial interval (ν) and the magnitude of R . Lines indicate combinations of ν and the fraction of ν in the latent period (f) that yield constant values of R . These R estimates assume a linearized SEIR model and the median extreme exponential growth rate.

Rapid application of control measures for a 1918-like influenza pandemic will be crucial. The short time between primary and secondary cases (ν) means that case numbers in a 1918-like pandemic will rise rapidly, with incident infections and the ensuing deaths doubling about every three days. Increased passenger travel relative to 1918 will facilitate the spread of a new virus across the globe. It is imperative that real-time surveillance information be shared freely, and that preventive measures be taken very early in a new pandemic. Therefore, while the relatively modest reproductive number estimated for 1918 pandemic influenza suggests the feasibility of controlling a similar future pandemic, significant planning and investment will be required to facilitate a rapid and effective response. □

Received 17 August; accepted 27 September 2004; doi:10.1038/nature03063.

- Patterson, K. D. & Pyle, G. F. The geography and mortality of the 1918 influenza pandemic. *Bull. Hist. Med.* **65**, 4–21 (1991).
- Anderson, R. M. & May, R. M. *Infectious Diseases of Humans: Dynamics and Control* (Oxford Univ. Press, Oxford, 1991).
- Webby, R. J. & Webster, R. G. Are we ready for pandemic influenza? *Science* **302**, 1519–1522 (2003).
- Nicholson, K., Webster, R. G. & Hay, A. J. *Textbook of Influenza* (Blackwell Science, Malden, Massachusetts, 1998).
- Webster, R. G., Bean, W. J., Gorman, O. T., Chambers, T. M. & Kawaoka, Y. Evolution and ecology of influenza A viruses. *Microbiol. Rev.* **56**, 152–179 (1992).
- Simonsen, L. *et al.* Pandemic versus epidemic influenza mortality: a pattern of changing age distribution. *J. Infect. Dis.* **178**, 53–60 (1998).
- Frost, W. H. Statistics of influenza morbidity with special reference to certain factors in case incidence and case fatality. *Public Health Rep.* **35**, 584–597 (1920).
- Longini, I. M. Jr, Halloran, M. E., Nizam, A. & Yang, Y. Containing pandemic influenza with antiviral agents. *Am. J. Epidemiol.* **159**, 623–633 (2004).
- Gog, J. R., Rimmelzwaan, G. F., Osterhaus, A. D. & Grenfell, B. T. Population dynamics of rapid fixation in cytotoxic T lymphocyte escape mutants of influenza A. *Proc. Natl Acad. Sci. USA* **100**, 11143–11147 (2003).
- Spicer, C. C. & Lawrence, C. J. Epidemic influenza in Greater London. *J. Hyg. (Lond.)* **93**, 105–112 (1984).
- Rvachev, L. A. & Longini, I. M. Jr A mathematical model for the global spread of influenza. *Math. Biosci.* **75**, 3–22 (1985).
- Fraser, C., Riley, S., Anderson, R. M. & Ferguson, N. M. Factors that make an infectious disease outbreak controllable. *Proc. Natl Acad. Sci. USA* **101**, 6146–6151 (2004).
- Jordan, E. O. *Epidemic Influenza; a Survey* (American Medical Assn, Chicago, 1927).
- Wolbach, S. B. Comments on the pathology and bacteriology of fatal influenza cases, as observed at Camp Devens, Mass. *Johns Hopkins Hosp. Bull.* **338**, 104–109 (1919).
- MacCallum, W. G. *Johns Hopkins Hospital Reports* (Johns Hopkins Press, Baltimore, 1921).
- Klotz, O. *Studies on Epidemic Influenza* 207–293 (Pittsburgh Univ. School of Medicine, Pittsburgh, 1919).
- Collins, S. D., Frost, W. H., Gover, M. & Sydenstricker, E. Mortality from Influenza and Pneumonia in 50 Large Cities of the United States 1910–1929. *Public Health Rep.* **45**, 2277–2328 (1930).
- Bureau of the Census. *Mortality Rates 1910–1920 with Population of the Federal Censuses of 1910 and 1920 and Intercensal Estimates of Population* (Govt Print. Off., Washington, 1923).
- Bureau of the Census. Population of the 100 largest urban places: 1920. (<http://www.census.gov/population/documentation/twps0027/tab15.txt>) (1998).
- Bureau of the Census. Population of the 100 largest urban places: 1910. (<http://www.census.gov/population/documentation/twps0027/tab14.txt>) (1998).
- Pearl, R. Influenza studies: further data on the correlation of explosiveness of outbreak of the 1918 epidemic. *Public Health Rep.* **36**, 273–298 (1921).
- Frost, W. H. The epidemiology of influenza. *J. Am. Med. Assoc.* **73**, 313–318 (1919).
- Olson, D. R., Simonsen, L., Edelson, P. J. & Morse, S. S. 4th Int. Conf. on Emerging Infectious Diseases [Abstract] (CDC, Atlanta, 2004).
- Stanley, L. L. Influenza at San Quentin Prison, California. *Public Health Rep.* **34**, 996–1008 (1919).
- Hill, A. N. & Longini, I. M. Jr The critical vaccination fraction for heterogeneous epidemic models. *Math. Biosci.* **181**, 85–106 (2003).
- Riley, S. *et al.* Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. *Science* **300**, 1961–1966 (2003).
- Lipsitch, M. *et al.* Transmission dynamics and control of severe acute respiratory syndrome. *Science* **300**, 1966–1970 (2003).
- Wallinga, J. & Teunis, P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am. J. Epidemiol.* **160**, 509–516 (2004).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. Wallinga and D. Olson for discussions and C. Raviola for data procurement. This work was supported in part by the National Defense Science and Engineering Graduate Fellowship (C.E.M.), the Medical Scientist Training Program Fellowship (C.E.M.), NIAID (J.M.R. and M.L.) and the Ellison Medical Foundation (M.L.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.E.M. (cmills@mit.edu).

Trail geometry gives polarity to ant foraging networks

Duncan E. Jackson¹, Mike Holcombe¹ & Francis L. W. Ratnieks²

¹Department of Computer Science, ²Department of Animal and Plant Sciences, University of Sheffield, Sheffield S1 4DP, UK

Pheromone trails are used by many ants to guide foragers between nest and food^{1–4}. But how does a forager that has become displaced from a trail know which way to go on rejoining the trail? A laden forager, for example, should walk towards the nest. Polarized trails would enable ants to choose the appropriate direction, thereby saving time and reducing predation risk. However, previous research has found no evidence that ants can detect polarity from the pheromone trail alone^{3,5–7}. Pharaoh's ants (*Monomorium pharaonis*) produce elaborate trail networks throughout their foraging environment⁸. Here we show that by using information from the geometry of trail bifurcations within this network, foragers joining a trail can adaptively reorientate themselves if they initially walk in the wrong direction. The frequency of correct reorientations is maximized when the trail bifurcation angle is approximately 60 degrees, as found in natural networks. These are the first data to demonstrate how ant trails can themselves provide polarity information. They also demonstrate previously unsuspected sophistication in the organization and information content of networks in insect societies.

Ants use chemical, visual and even magnetic cues to orient themselves correctly in their foraging environment^{4,9}. Some pheromone-trail-following ants readily correct their course, when displaced, by using their memory of environmental cues (landmarks and sun compass)^{3,10,11}. For example, in the common garden ant *Lasius niger*^{10,11} movement of visual cues by the experimenter results in course adjustments by the ant, while in the Argentine ant *Linepithema humilis*¹¹ and *La. fuliginosus*³ pheromone trails are always followed, regardless of whether visual cues are moved. Thus ants using external orientation cues can readily determine direction when on a trail. However, species that are principally guided by the chemical trail itself may find reorientation harder. One solution would be to follow the trail to its conclusion at nest or food, and then reorientate if necessary. But making such long trips to determine directionality would be costly. A better solution would be to have polarized trails.

Previous experimental studies, however, found no evidence of intrinsic polarity in pheromone trails^{3,6,7,10}. These studies assumed that polarity information contained in trails must take the form of a pheromone concentration gradient. However, pheromone concentration is not the only mechanism that could polarize trails. In particular, trail geometry could supply this information. Ant foraging trails are often thought of as a single line, but in fact most trail-laying ants produce complex trail networks branching throughout their foraging environment^{8,12–15}, suggesting that network geometry could provide polarity information. Acosta *et al.*¹⁶ studied networks of three species of leafcutter ants (*Atta sexdens*, *A. capiguara* and *A. laevigata*) and one seed harvester ant (*Messor barbarus*), and found that they shared one common geometrical feature—the mean angle between trail bifurcations as they branch out from the nest is 50–60°.

Our results show that Pharaoh's ant colonies (*Mo. pharaonis*) produce similar networks (Fig. 1a–c) with a mean bifurcation angle of 53° (Fig. 1c). Pharaoh's ants readily form trail networks, even before discovering food, which highlights their reliance on pheromone trails for orientation^{8,17}.

Information inherent in trail bifurcations could be used by

foragers to determine trail polarity. For example, ants encountering a 60° bifurcation while travelling away from the nest would have a choice of two paths, each deviating by approximately 30° from their current heading. Returning foragers would also have a choice of two paths but, in contrast to outbound ants, one deviates by only a small amount (30°) from the current heading while the other deviates greatly (120°). The trail making a large deviation would lead away from the nest whereas the one with a small deviation only would lead to the nest.

In experiment 1, we determined whether individual Pharaoh's ant foragers could reorientate when placed onto a natural and complete trail network lacking other ants, if they were initially walking the wrong way. Our results clearly show that they can. Most of the foragers (70%) reoriented themselves (75%, $n = 20$, of fed foragers; and 65%, $n = 20$, of unfed foragers), which is highly significant when compared to the 12.5% ($n = 40$) that reoriented in the control group (χ^2 test, degrees of freedom, d.f. = 1, $P = < 0.001$). Corrections to routes took the form of reorientations (U-turns or large deviations) at a trail bifurcation (53.6%, $n = 28$) or U-turns following a bifurcation (46.4%, $n = 28$), which suggested that bifurcations are important.

In experiment 2, we obtained strong support for the geometry hypothesis. In this experiment we allowed individual ants (fed or unfed) to walk along experimental straight or bifurcating trails. Tests conducted on straight trails showed that any reorientations occurring were as likely to be correct (R_c) as incorrect in relation to the polarity of the trail when it was originally formed (R_i), that is, 7% versus 6% ($R_c/R_i = 1.17$). These course changes were all U-turns and occurred at a low frequency (6.5%, $n = 200$). In addition, 17.5% of ants left the trail altogether. In contrast, a single trail bifurcation (55°) facilitated many more course corrections, and so

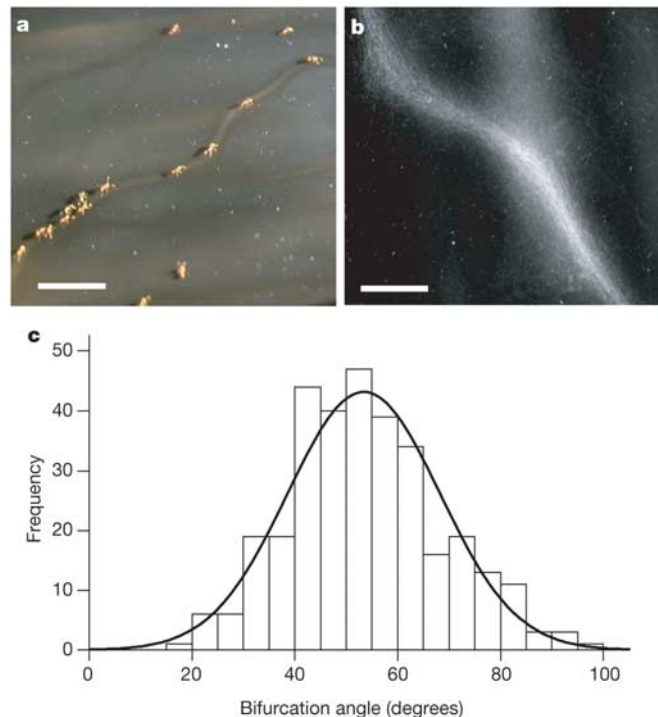


Figure 1 Pheromone trail networks of Pharaoh's ants on a smoked glass surface. **a**, Part of a network showing bifurcations to smaller trails (scale bar, 1 cm). **b**, Close-up of a single bifurcation (scale bar, 0.5 cm). **c**, Analysis of bifurcation angles from four separate trail networks: mean angle = 53.4°, s.d. = 14.8°, $n = 321$; mean distance between bifurcations = 2.9 cm, s.d. = 2.3 cm, $n = 485$. Solid line, normal distribution curve.

presumably provided much more information. 43% of fed ants ($n = 100$) made U-turns upon meeting the bifurcation point when walking in the 'wrong' direction: that is, away from the nest (Supplementary Table 1). Conversely, only 8% of fed ants walking the 'correct' way made U-turns or other corrections that led to them incorrectly heading away from the nest. This is similar to the 'background' U-turn frequency observed on straight trails (6.5%). Data from unfed ants are very similar. 47% ($n = 100$) made course corrections at the bifurcation point when moving the 'wrong' way: that is, towards the nest. Again, only 8% walking the 'correct' way made incorrect course changes. Overall, correct course changes (R_c) occurred far more frequently than incorrect ones (R_i), producing a correct to incorrect ratio ($R_c:R_i$) of 5.63 (45%:8%). Note that although only 45% of ants corrected their orientation at a single bifurcation, real networks contain many bifurcations and many opportunities for course correction.

To control for the possibility that alternative orientation cues might confound our results, these experiments were repeated in two alternative positions, by rotating the test trail sections through 90° and 180° relative to their original position when established. The results were not significantly different to those obtained in the original position. (For the straight trail: 90° rotation, χ^2 test, d.f. = 3, $P = 0.054$, $R_c:R_i = 0.6$; 180° rotation, χ^2 test, d.f. = 3, $P = 0.064$, $R_c:R_i = 1.17$. For the bifurcating trail: 90° rotation, $\chi^2 = 1.185$, d.f. = 3, $P = 0.553$, $R_c:R_i = 6.33$; 180° rotation, $\chi^2 = 4.565$, d.f. = 3, $P = 0.102$, $R_c:R_i = 4.06$.) These results confirm that Pharaoh's ants making course corrections while walking on trails use pheromone trails for orientation, rather than external cues. They also confirm that straight sections of the trails are not polarized.

The above trials used experimental trail bifurcations of 55° to investigate course corrections at the natural angle of trail bifurcations. However, it was also important to investigate reorientation at additional bifurcation angles. In particular, the trail-geometry hypothesis predicts that the ability of ants to correctly reorientate will diminish as the angle approaches 120°. This is because a bifurcation of 120° is symmetrical and conveys no polarity information. Using the same methods as above, we studied angles of 30, 45, 60, 75, 90, 105 and 120 degrees. Results for correct and incorrect reorientations are presented in Fig. 3a and reorientation ratios ($R_c:R_i$) in Fig. 3b. The relationship is highly nonlinear

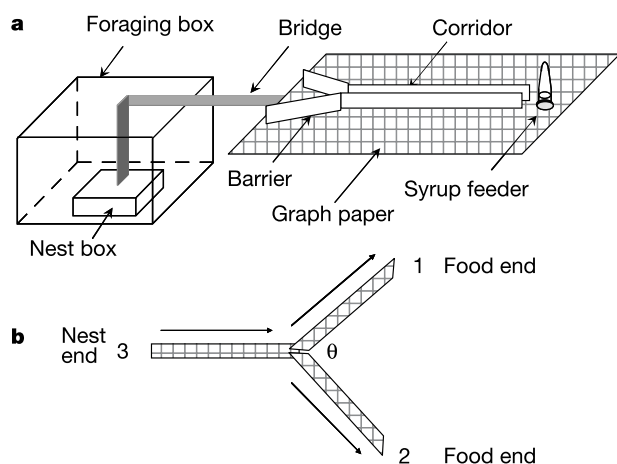


Figure 2 Experimental set-up used to form straight trails and assembly of trail bifurcations. Foraging ant traffic was constrained to produce straight pheromone trails on paper (a), which were cut into sections (b) and reassembled to form the requisite trail bifurcations with variable angle $\theta = 0$ –120°. Start locations were designated as 3, the nest direction (left) and 1 or 2, the food directions (right). Arrows show that the original direction to food on the trail was preserved when bifurcations were prepared.

($r^2 = 0.016$; analysis of variance, ANOVA, d.f. = 8, $F = 0.11$, $P = 0.749$) but fits with high statistical significance to a quadratic curve ($R^2 = 0.923$; ANOVA, d.f. = 8, $F = 36.16$, $P < 0.001$). Bifurcation angles from 30–90° were highly effective, with $R_c:R_i$ ratios ranging from 3.28 at 30° to a maximum at 60° ($R_c:R_i = 5.8$), then declining to 4.26 at 90°. As predicted, the ratio of 1.03 at 120° was not significantly different to that of a straight trail (1.17), demonstrating an absence of polarity information at 120° (χ^2 test, d.f. = 1, $P = 0.952$). Angles close to 60° also minimized the proportion of ants leaving the trail. The quadratic curve fitted to trail-leaving frequency ($R^2 = 0.729$; ANOVA, d.f. = 8, $F = 8.71$, $P = 0.019$) showed (Fig. 3c) that angles in the range 33–97° had the lowest leaving rates (14–15%).

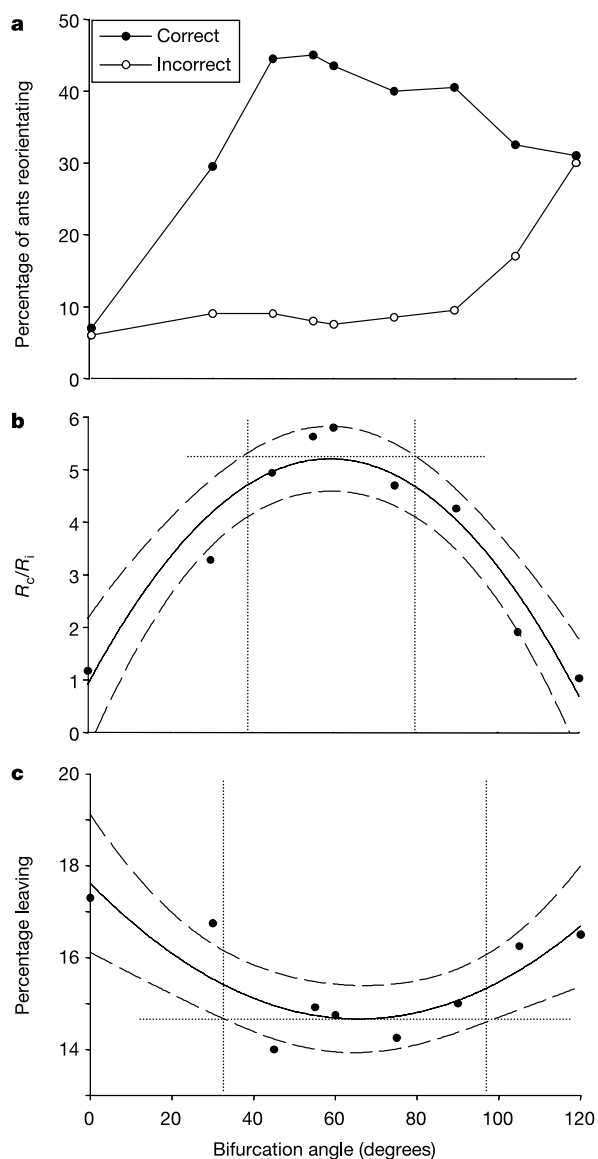


Figure 3 Outcomes of individual trail following tests using straight (0°) and bifurcating (30–120°) trail sections. **a**, Percentage of foragers making correct and incorrect reorientations. For each point $n = 200$, except at 0° where $n = 100$. **b**, Ratio of correct to incorrect reorientations ($y = -0.0012x^2 + 0.1449x + 0.9175$) with 95% confidence interval (dashed lines, 36–81°). **c**, Percentage of individual ants departing from test trails at each angle where $n = 400$ for angles 30–120° and $n = 200$ for 0° ($y = -0.0007x^2 + 0.0933x + 17.625$) with 95% confidence interval (dashed lines, 33–97°). (Dotted lines show the 95% confidence intervals for the angles with the maximum ratio, in both **b** and **c**.)

The branching geometry of Pharaoh's ant trails enables individual foragers moving in the wrong direction to reorient themselves. With bifurcations close to the natural angle the effect is strong both in the ratio of correct to incorrect reorientations (about 5:1) and in the proportion of ants who reorient themselves at a single bifurcation (~45%). Our results also support the geometric prediction that straight trails and those bifurcating at 120° do not result in adaptive course corrections.

The physicist Feynman speculated that a system of pheromones laid in a linear pattern of A-B-space-A-B-space might provide polarity information¹⁸. Realistically, however, such a system could only work for trails laid by one ant, because the message would be lost when trails were overlaid. Pheromone gradients could, in principle, provide polarity information but seem not to be used by ants, perhaps because they are unreliable. Trails nearer the nest should have higher pheromone concentrations caused by greater traffic. But the reverse can be true, particularly when a food source is newly discovered. In fact, the trail bifurcation system that we have discovered is reliable and much simpler.

Our results show that bifurcation angles similar to those found in natural networks (~60°) give the best reorientation ratios. Other natural transport and communication systems with branching, such as plant roots or cardiovascular systems, are thought to have evolved to minimize energy expenditure in resource distribution^{19–23}. Similarly, trail bifurcations of approximately 60° may actually optimize the flow of ants through the network, particularly in a natural situation where traffic is bi-directional²⁴. Thus, trail networks could have been evolutionarily optimized to achieve an efficient flow of resources to the nest and ants may exploit this emergent property of networks in orientation. The geometry of these trail networks constructed by Pharaoh's ants provides a subtle and previously unsuspected source of information, enabling the colony to forage more efficiently. □

Methods

Study species

Study colonies of Pharaoh's ants (*Mo. pharaonis*) contained 1,200–2,500 workers, brood at all stages and multiple queens (12–50), and were housed in wooden nest boxes (11 cm × 8 cm × 2 cm high) held within a large plastic foraging box (45 cm × 30 cm × 15 cm high) in a climate-controlled room (24 ± 2 °C, relative humidity = 30%, light/dark ratio 12 h:12 h). *Mo. pharaonis* has monomorphic workers, ~2 mm body length, and nest-mate recognition is absent in this unicolonial ant⁴. Frequent splitting and combining of the colonies made the six study colonies genetically similar. Colonies were given fresh water ad libitum and fed daily with mealworm (*Tenebrio*) larvae, weekly with apple sauce, and monthly with dried egg yolk.

Quantifying the geometry of trail networks formed on smoked glass

Sheets of toughened glass (39.2 cm × 27.6 cm) were held over a wax candle flame to coat them with a fine layer of soot. A colony was then given access to one edge of a smoked sheet via a drawbridge. A syrup feeder was placed at the opposite edge. Colonies were allowed to forage for 2 h, which was sufficient time to establish trail networks covering the glass sheet. The ants were shaken off and the trail networks, visible as lines where the soot had been worn away (Fig. 1), were traced onto transparent plastic and the bifurcation angles measured.

Individual forager reorientation on smoked glass trail networks (experiment 1)

In each orientation trial individual forager ants were collected from the floor of their colony's foraging box and put into one of two internal states, 'unfed' or 'fed', as follows. A group of 25 ants were held for 2–2.5 h in an empty plastic box (unfed), or in a box containing three 1 M sucrose solution feeders. Of the ants with access to food only those with "a visibly enlarged gaster, which appeared striped because all the abdomen was distended, exposing the inter-segmental membranes"²⁵ were defined as fed. We placed single fed or unfed ants on a natural trail network (smoked glass, as above) from which the ants had been removed within the previous 10 min. Each ant was oriented on a trail facing the 'wrong' direction (that is, away from the nest) when fed and facing towards the nest when unfed. Twenty ants in each of the two states were observed until they reached either the feeder or the drawbridge via the trail network. We noted how many fed and unfed ants correctly reoriented on the trail network and where they did this relative to trail bifurcations. Control samples of 20 ants in each of the two states were also placed on trails facing the 'correct' direction and their destinations recorded.

Individual trail bifurcation tests (experiment 2)

To construct experimental bifurcations we constrained the foraging of *M. pharaonis*

colonies using plastic barriers to form straight pheromone trails on a piece of paper leading from the nest to a syrup feeder. We temporarily linked a colony box to a foraging arena (140 cm × 70 cm) with a drawbridge. Part of the arena was covered with two A4 (29.5 cm × 21.0 cm) sheets of ECF (eucalyptus chlorine-free) paper photocopied with a 2-mm grid. Two polycarbonate strips (60 cm × 4 cm × 0.5 cm thick), coated with Fluon to prevent ants from climbing them, were placed on the paper to form a 4-mm-wide corridor leading to the syrup feeder, which was held within a section of Fluon-coated plastic pipe (8 cm in diameter). A total of 2,000 ants were counted passing the mid-point of the corridor (both directions combined) before ants were shaken from the paper. The 60 cm × 0.4 cm section along which the ants had walked and deposited trail pheromone was cut from the paper. Experimental bifurcations were constructed (except for 0°, which was a 30 cm straight trail) using three 15-cm trail sections, by cutting their tips to form the appropriate angle (30°, 45°, 55°, 60°, 75°, 90°, 105°, 120°), then securing the cut pieces of paper with adhesive tape to a fresh sheet of backing paper (Fig. 2). Because trails decay rapidly²⁶ we used trial bifurcations for a maximum of 20 min after the ants forming the trail had been removed from the paper. For each angle we tested 100 individual ants starting at the nest end (position 3) and 50 ants at each of the two food ends (positions 1 and 2), for each state (unfed and fed, as detailed in experiment 1). For straight trails we used only 50 ants in each state at each end. Trail decay limited separate trials to 20–50 ants on each occasion so data from several trials were pooled per bifurcation angle. To control for visual and magnetic cues, we repeated experiments for straight trails and 55° bifurcations but with the trail set-up rotated through 90° and 180° relative to the standard position.

Received 1 July; accepted 6 October 2004; doi:10.1038/nature03105.

- Schöne, H. *Spatial Orientation* (Princeton Univ. Press, Princeton, 1984).
- Camazine, S. et al. *Self-organisation in Biological Systems* 217–255 (Princeton Univ. Press, New Jersey, 2001).
- Carthy, J. D. The orientation of two allied species of British ant. II. Odour trail laying and following in *Acanthomyops (Lasius) fuliginosus*. *Behaviour* **3**, 304–318 (1951).
- Hölldobler, B. & Wilson, E. O. *The Ants* (Belknap Press, Harvard Univ., Cambridge, Massachusetts, 1990).
- Carthy, J. D. Odour trails of *Acanthomyops fuliginosus*. *Nature* **166**, 154 (1950).
- Sudd, J. H. An *Introduction to the Behaviour of Ants* Ch. 6 (Edward Arnold, London, 1967).
- Wilson, E. O. Chemical communication among workers of the fire ant, *Solenopsis saevissima*. I. The organisation of mass foraging. *Anim. Behav.* **10**, 134–164 (1962).
- Sudd, J. H. The foraging method of Pharaoh's ant, *Monomorium pharaonis* (L.). *Anim. Behav.* **8**, 67–75 (1960).
- Banks, A. N. & Syrgley, R. B. Orientation by magnetic field in leaf-cutter ants, *Atta colombica* (Hymenoptera: Formicidae). *Ethology* **109**, 835–846 (2003).
- Carthy, J. D. The orientation of two allied species of British ant. I. Visual direction finding in *Acanthomyops (Lasius) niger*. *Behaviour* **3**, 275–303 (1951).
- Aron, S., Beckers, R., Deneubourg, J. & Pasteels, J. M. Memory and chemical communication in the orientation of two mass-recruiting ant species. *Insectes Soc.* **40**, 369–380 (1993).
- Fowler, H. G. Foraging trails of leaf-cutting ants. *Proc. NY Entomol. Soc.* **86**, 132–136 (1978).
- Hölldobler, B. Home range orientation and territoriality in harvester ants. *Proc. Natl Acad. Sci. USA* **71**, 3274–3277 (1974).
- Quinet, Y. & Pasteels, J. M. Spatiotemporal evolution of the trail network in *Lasius fuliginosus* (Hymenoptera, Formicidae). *Belg. J. Zool.* **121**, 55–72 (1991).
- Rettenmeyer, C. W. Behavioral studies of army ants. *Univ. Kansas Sci. Bull.* **44**, 281–465 (1963).
- Acosta, F. J., Lopez, F. & Serrano, J. M. Branching angles of ant trunk trails as an optimization cue. *J. Theor. Biol.* **160**, 297–310 (1993).
- Fourcassie, V. & Deneubourg, J. L. The dynamics of collective exploration and trail-formation in *Monomorium pharaonis*: experiments and model. *Physiol. Entomol.* **19**, 291–300 (1994).
- Feynman, R. *Surely You Are Joking, Mr Feynman?* (Norton, New York, 1985).
- Barker, S. B., Cumming, G. & Horsfield, K. Quantitative morphometry of the branching structure of trees. *J. Theor. Biol.* **40**, 33–43 (1973).
- Leopold, L. B. Trees and streams: the efficiency of branching patterns. *J. Theor. Biol.* **31**, 339–354 (1971).
- Thompson, D. W. *On Growth and Form* (Cambridge Univ. Press, Cambridge, 1917).
- West, G. B., Brown, J. H. & Enquist, B. J. A general model for the origin of allometric scaling laws in biology. *Science* **276**, 122–126 (1997).
- West, G. B., Brown, J. H. & Enquist, B. J. A general model for the structure and allometry of plant vascular systems. *Nature* **400**, 664–667 (1999).
- Dussutour, A., Fourcassie, V., Helbing, D. & Deneubourg, J. Optimal traffic organization in ants under crowded conditions. *Nature* **428**, 70–73 (2004).
- Tennant, L. E. & Porter, S. D. Comparison of diets of two fire ant species: solid and liquid components. *J. Entomol. Sci.* **26**, 450–465 (1991).
- Jeanson, R., Ratnieks, F. L. W. & Deneubourg, J. L. Pheromone trail decay rates on different substrates in the Pharaoh's ant, *Monomorium pharaonis*. *Physiol. Entomol.* **28**, 192–198 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank K. Boomsma, A. Bourke, V. Fourcassie, A. Hart, L. Keller, S. Martin and G. Theraulaz for discussions and comments on the manuscript. We also thank P. Mitchell for statistical advice. This work was funded by the EPSRC and British Telecom.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.J. (duncan@dcs.sheffield.ac.uk).

Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment

Mary Ann Moran¹, Alison Buchan³, José M. González⁴, John F. Heidelberg⁵, William B. Whitman², Ronald P. Kiene⁶, James R. Henriksen², Gary M. King⁷, Robert Belas⁸, Clay Fuqua⁹, Lauren Brinkac⁵, Matt Lewis⁵, Shivani Johri⁵, Bruce Weaver⁵, Grace Pai⁵, Jonathan A. Eisen⁵, Elisha Rahe⁹, Wade M. Sheldon¹, Wenying Ye¹, Todd R. Miller⁸, Jane Carlton^{5,10}, David A. Rasko⁷, Ian T. Paulsen⁵, Qinghu Ren⁵, Sean C. Daugherty⁵, Robert T. Deboy⁵, Robert J. Dodson⁵, A. Scott Durkin⁵, Ramana Madupu⁵, William C. Nelson⁵, Steven A. Sullivan⁵, M. J. Rosovitz⁷, Daniel H. Haft⁵, Jeremy Selengut⁵ & Naomi Ward^{5,8}

¹Department of Marine Sciences and ²Department of Microbiology, University of Georgia, Athens, Georgia 30602, USA

³Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, Connecticut 06520, USA

⁴Departamento de Microbiología y Biología Celular, Universidad de La Laguna, 38206 La Laguna, Tenerife, Spain

⁵The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA

⁶Department of Marine Sciences, University of South Alabama, Mobile, Alabama 36688, USA and Dauphin Island Sea Lab, Dauphin Island, Alabama 36528, USA

⁷Department of Biochemistry, Microbiology and Molecular Biology, Darling Marine Center, University of Maine, Walpole, Maine 04573, USA

⁸Center of Marine Biotechnology, University of Maryland Biotechnology Institute, 701 East Pratt Street Baltimore, Maryland 21202, USA

⁹Department of Biology, 1001 E. 3rd Street, Jordan Hall 418, Indiana University, Bloomington, Indiana 47405, USA

¹⁰Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, Maryland 21205, USA

Since the recognition of prokaryotes as essential components of the oceanic food web¹, bacterioplankton have been acknowledged as catalysts of most major biogeochemical processes in the sea. Studying heterotrophic bacterioplankton has been challenging, however, as most major clades have never been cultured² or have only been grown to low densities in sea water^{3,4}. Here we describe the genome sequence of *Silicibacter pomeroyi*, a member of the marine Roseobacter clade (Fig. 1), the relatives of which comprise ~10–20% of coastal and oceanic mixed-layer bacterioplankton^{2,5,6,7}. This first genome sequence from any major heterotrophic clade consists of a chromosome (4,109,442 base pairs) and megaplasmid (491,611 base pairs). Genome analysis indicates that this organism relies upon a lithoheterotrophic strategy that uses inorganic compounds (carbon monoxide and sulphide) to supplement heterotrophy. *Silicibacter pomeroyi* also has genes advantageous for associations with plankton and suspended particles, including genes for uptake of algal-derived compounds, use of metabolites from reducing microzones, rapid growth and cell-density-dependent regulation. This bacterium has a physiology distinct from that of marine oligotrophs, adding a new strategy to the recognized repertoire for coping with a nutrient-poor ocean.

The genome sequence of the coastal bacterioplankton *S. pomeroyi* DSS-3 was determined by the random whole-genome shotgun method and found to contain 4,283 predicted coding sequences (CDS, Table 1). The megaplasmid contains 10.4% of the CDS, with biases towards genes involved in energy metabolism, transport and regulation (see Supplementary Table S1). Compared with other α -Proteobacteria for which complete genome sequences are available, *S. pomeroyi* has the highest proportion of genes coding for signal transduction (1.6% versus $0.23 \pm 0.39\%$ ($\pm$ s.d.) for 14 other

α -Proteobacteria) and transport/binding proteins (12.1% versus $8.1 \pm 3.0\%$) (see Supplementary Fig. S1), potentially reflecting an enhanced ability to sense and respond to conditions outside the cell.

Two operons encoding aerobic carbon monoxide dehydrogenases for the oxidation of CO to CO₂ are present in the *S. pomeroyi* genome (*coxSML*). The absence of ribulose biphosphate carboxylase or other complete pathways for autotrophy suggests that the bacterium gains energy, but not carbon, through carboxidotrophy. As all previously characterized carboxidotrophs are autotrophic⁸, CO oxidation was experimentally confirmed in *S. pomeroyi* and found to occur at CO concentrations typical of surface sea water (10 nM in coastal and 2 nM in oceanic regions) (see Supplementary Fig. S2). *Silicibacter*-like lithoheterotrophs might act as a microbial CO sink in the surface ocean—this poorly characterized process consumes 10–60 teragrams of C (as CO) annually and buffers the partial pressure of this greenhouse gas in the ocean⁹. A cluster encoding reduced inorganic sulphur oxidation (*soxRSVWXY-ZABCD*) provides another mechanism for lithoheterotrophic growth in *Silicibacter*. The presence of a reduced inorganic sulphur compound was experimentally confirmed to enhance biomass production in acetate-grown *S. pomeroyi* cultures by 45% relative to cultures receiving no sulphur (see Supplementary Fig. S3). Analysis of codon usage in the *S. pomeroyi* genome indicates that both *coxL* and *soxB* have codon adaptive indices (CAIs) higher than >90% of the genes (0.36 and 0.35, respectively). CAIs have been found to correlate positively with levels of transcription in other genomes.

Because cultured bacteria might not be appropriate analogues of their uncultured relatives, the importance of lithoheterotrophy among marine bacterioplankton was assessed by an analysis of gene stoichiometry in the Sargasso Sea environmental shotgun library¹⁰. *coxL* was present in the library at an abundance equivalent to one per 14 bacterial cells and *soxB* at one per 10 cells (Table 2). As the Roseobacter clade accounted for only 3% of Sargasso 16S ribosomal RNA genes (see Supplementary Table S2), lithoheterotrophic strategies involving CO and inorganic sulphur compounds probably have distribution beyond this taxonomic group. Both *coxL* and *soxB* are also represented in a smaller coastal environmental genomics library¹¹. CO is ubiquitous in marine surface waters owing to photo-oxidation of dissolved organic matter (DOM)⁹, and sulphide has been measured in reducing microzones of marine snow¹² and might be generated from the degradation of organic sulphur compounds. Energy derived from oxidation of reduced inorganic compounds could allow greater heterotrophic efficiency for marine bacterioplankton by decreasing the fraction of assimilated organic carbon respired and hence increasing bacterial growth yield. Alternatively, lithotrophy could enhance anaplerotic CO₂

Table 1 General features of the *S. pomeroyi* genome

Feature		
Total number of coding sequences	4,283	
Number of rRNA operons (16S, 23S, 5S)	3	
Number of tRNA genes	53	
Number of structural RNA genes	2	
Proteins similar to proteins of known function and role (% of total proteins)	59.7	
Proteins of unknown function (% of total proteins)	16.8	
Conserved hypothetical proteins (% of total proteins)	15.3	
Hypothetical proteins (% of total proteins)	8.2	
	Main chromosome	Megaplasmid
Molecule length (bp)	4,109,442	491,611
G+C content (%)	64.2	62.8
Number of coding sequences	3,838	445
Average size of coding sequences (bp)	966	997
Percentage of coding sequence	90.2	90.3

fixation via pyruvate carboxylase, as has been found for *Sulfitobacter pontiacus*—a member of the Roseobacter clade—in the presence of thiosulphate¹³. Either alternative would influence the pathways and efficiencies of carbon flow through the marine microbial loop. Along with recent discoveries of bacterial photoheterotrophy in marine surface waters^{14,15}, they imply that mixed metabolic schemes might be particularly successful in the ocean.

High numbers of peptide transporters (16 ATP-binding cassette (ABC)-type systems), branched chain amino acid transporters (ten ABC-type systems) and amino acid efflux proteins (seven RhtB family systems) suggest that proteins are an important carbon source for *Silicibacter*. Six ABC-type transporter systems for putrescine and spermidine are present (no other sequenced genome has more than three); these compounds are produced in marine phytoplankton and zooplankton to regulate cell proliferation and bloom formation¹⁶. Transport systems for algal osmolytes feature significantly in the *S. pomeroyi* genome, and include five systems for transporting glycine betaine and/or dimethylsulphoniopropionate (DMSP) (OpuA and OpuD). Only the marine bacterium *Oceanobacillus iheyensis* has more such systems (with 16, see Supplementary Table S3). There is also a transport system and degradation pathway for taurine (*tauABCR*, *xsc*, *pta*, *tpa*). Organic matter from living plankton and detrital particles in the surface mixed layer is composed of $\geq 50\%$ protein¹⁷, and osmolytes such as DMSP can comprise 20% of the cytoplasm of some marine phytoplankton¹⁸. Several genome features suggest that organic matter derived from plankton and marine snow are important substrates for *Silicibacter*-like organisms. Members of the Roseobacter clade are commonly found in association with algal cells¹⁹ and have a depth distribution that matches that of phytoplankton⁵. *S. pomeroyi* is motile and is

potentially able to position itself in favourable microniches in association with plankton and particles (31 genes encode elements for motility via a polar complex flagellum²⁰). However, there are no open reading frames (ORFs) with strong homology to known chemotaxis proteins or methyl-accepting chemotaxis transducers, suggesting either that the organism is motile but not chemotactic or that chemotaxis occurs by an unknown mechanism, either of which is unusual.

Analysis of flanking genes of an uncommonly high number of TRAP transporter systems (26 systems; no other sequenced genome has more) indicates an ability of *S. pomeroyi* to transport carboxylic acids of the type produced in surface waters during photo-oxidation of DOM²¹, including glyoxylate and acetate. Other secondary transporters target malonate and formate. Carboxylic acids together with CO form a pool of biologically labile DOM photoproducts that can support 20% of bacterioplankton biomass production in near-surface coastal waters at temperate latitudes²¹. While it is anticipated that many marine bacteria assimilate selected photoproducts, *Silicibacter*-like organisms are genetically positioned to make thorough use of the mixture of compounds formed by DOM photo-oxidation.

Two putative quorum-sensing systems (*luxI/luxR* homologue pairs) are present on the *S. pomeroyi* chromosome. The presence of these *luxI* homologues is consistent with detection of at least three acylhomoserine lactone (AHL)-type signal molecules in culture extracts and distinct AHL-type activities for each homologue when expressed in *Escherichia coli* (see Supplementary Fig. S4). The presence of two quorum-sensing systems is in contrast to what is expected for an oligotrophic marine bacterium with conservative metabolic strategies. Quorum sensing would provide

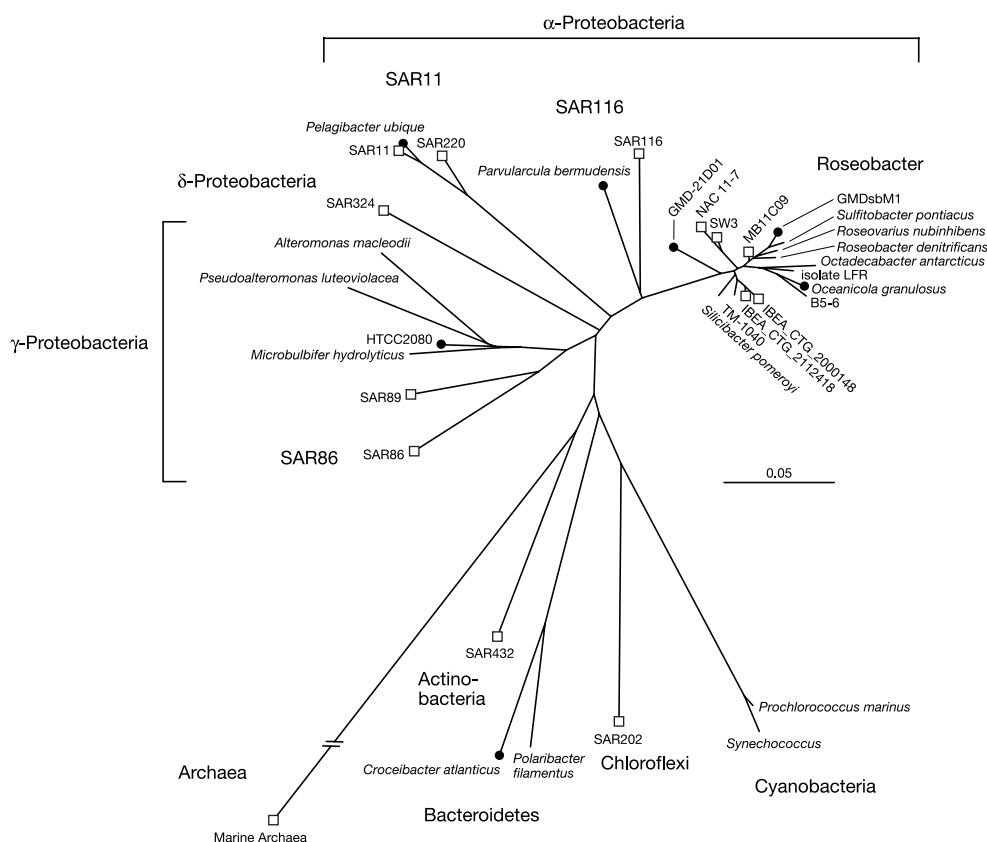


Figure 1 Phylogenetic tree of 16S rRNA gene sequences from the Roseobacter clade and other major marine taxa. Sequences include those from uncultured bacterioplankton

(open square) and from cultured bacterioplankton isolated at very low nutrient concentrations (filled circle). Scale bar shows Jukes–Cantor evolutionary distance.

Table 2 **Stoichiometry of selected genes in the Sargasso Sea shotgun library***

Query gene	Sample 1	Sample 2	Sample 3	Frequency in cells† Sample 4	Sample 5	Sample 6	Sample 7
16S rRNA	1 in 0.31	1 in 0.48	1 in 0.46	1 in 0.55	1 in 0.65	1 in 0.33	1 in 0.54
Roseo 16S rRNA	1 in 68	1 in 114	1 in 80	1 in 69	0	1 in 5	1 in 44
<i>coxL</i>	1 in 11	1 in 22	1 in 10	1 in 15	0	1 in 3	1 in 41
<i>soxB</i>	1 in 12	1 in 9	1 in 9	1 in 8	1 in 14	1 in 4	1 in 16
<i>nirS</i>	1 in 225	0	0	0	0	0	0

*Samples 1–7 from the unassembled Sargasso Sea shotgun library¹⁰.

†Calculated as: number of target genes divided by number of *recA* homologues for the target genes 16S rRNA, *coxL*, *soxB* and *nirS* (rows 1, 3, 4 and 5). For Roseobacter 16S rRNA genes (row 2), calculated as: (number of 16S rRNA genes × 0.47) divided by the number of *recA* homologues. This assumes that Roseobacter clade cells contain an average of 2.1 rRNA operons based on the 16S rRNA:*recA* gene ratio in the complete Sargasso data set (this table). *recA* is typically a single-copy gene. BLAST searches against trace files identified 225, 217, 228, 231, 28, 16 and 84 *recA* homologues in Sargasso samples 1–7, respectively.

regulatory flexibility to manage a complex suite of metabolic pathways, possibly associated with switching between particle-associated (high population density, high substrate availability) and free-living (low population density, low substrate availability) stages²². The presence of three rRNA operons in the *S. pomeroyi* genome is consistent with the potential to respond rapidly to changes in resource availability²³, and a pathway for polyhydroxyalkanoic acid synthesis (*phaZCP*, *phbAB*) provides a mechanism for carbon and energy storage when substrates are abundant.

Four transporters for ammonium (*amt*) and one for urea (*urtABC*) are present in the *S. pomeroyi* genome, as are genes for ammonium (*glnA*, *gdhA*) and urea (*ureABCDEF*) assimilation. Genes for assimilating nitrate and nitrite were not identified. Experiments confirmed that ammonium and urea are used as nitrogen sources, whereas nitrate and nitrite are not (see Supplementary Fig. S5). Nitrite is toxic at high concentrations (>500 µM), potentially because NO accumulates through the partial denitrification pathway (NO₂ to N₂) encoded on the *S. pomeroyi* megaplasmid (*norFEDQBC*, *nnrS*, *nirSECF*, *D-L* fusion, *GHJN*, *nosRZDFYL*). Like CO and other DOM photodegradation products, ammonium and urea are regenerated in marine surface waters. Thus, nutrient, energy and carbon acquisition strategies in *S. pomeroyi* coincide with Roseobacter depth distributions (highest abundance in surface waters)^{2,5}. Although a single cultured organism cannot represent the breadth of features embodied in a diverse taxon that is broadly distributed over distinct hydrographical regions of the ocean²⁴, the emerging picture of *S. pomeroyi* physiology fits distribution patterns previously observed for Roseobacter clade members.

This genome sequence from a major bacterioplankton clade reveals an organism equipped to take advantage of transient occurrences of high-nutrient niches within a bulk low-nutrient environment. Living and dead plankton and microscale 'hot spots' of the surface ocean²⁵ might provide such niches. Lithoheterotrophic growth could allow *Silicibacter*-like bacterioplankton to use a greater proportion of organic carbon for biomass production as it becomes available. Although most ecologically relevant marine heterotrophs were previously assumed to be oligotrophs that subsist on dilute organic substrates dissolved in sea water²⁶, an 'opportunistic' strategy might be a successful alternative. The available metagenomic data from coastal¹¹ and oceanic¹⁰ sites (Table 2) indicate that such a strategy is not atypical among marine bacterioplankton. The genome sequence of this cultured marine microbe provides a window into the ecological strategies that maintain a significant fraction of the prokaryotes in the ocean. □

Methods

Isolation of *S. pomeroyi* (ATCC700808)

Sea water from southeastern US coastal waters (salinity = 31) was diluted with filter-sterilized sea water and enriched with 10 µM DMSP for two weeks. A small white colony was selected from a low-nutrient seawater plate spread with 0.1 ml of the enriched sea water²⁰.

Sequencing and annotation

The complete genome sequence of *S. pomeroyi* was determined using the whole-genome shotgun method as described²⁷, with libraries of 1–2 kb and 12–15 kb. Closure of physical and sequencing gaps was performed by primer walking, sequencing of transposon-tagged libraries of large-insert clones and multiplex polymerase chain reaction (PCR). Assembly was performed with the TIGR Assembler²⁷, and repeats were identified using RepeatFinder. G+C skew and oligoskew analyses²⁷ identified a putative origin of replication, and base pair 1 was assigned adjacent to the glucose-inhibited division protein A gene (*gidA*).

An initial set of ORFs predicted to encode proteins was identified using GLIMMER. ORFs consisting of <30 codons or containing overlaps were eliminated, and frameshifts and point mutations were corrected or designated authentic. Functional assignment of genes, identification of membrane-spanning domains, determination of paralogous gene families and identification of regions of unusual nucleotide composition were performed²⁷. Phylogenomic analysis was used to assist with functional predictions²⁷, and comparative genome analyses were performed using the Comprehensive Microbial Resource (<http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl>). CAI index was calculated using the CAIC package²⁸.

Carbon monoxide oxidation

Three millilitres of culture (absorbance at 600 nm (*A*₆₀₀) of 1) were incubated in triplicate 160-ml serum bottles at 30 °C, shaking at 250 r.p.m. CO was added at 10 p.p.m. and uptake was monitored to lower limits of 0.13 ± 0.05 p.p.m. using headspace gas chromatographic (GC) analysis (see Supplementary Fig. S2).

Thiosulphate assay

One millilitre of overnight culture was added to 100 ml of basal medium plus acetate (2, 5 or 10 mM) with or without 10 mM thiosulphate. Cultures were incubated at 30 °C and shaken at 250 r.p.m. in the dark and *A*₆₀₀ was measured (see Supplementary Fig. S3).

Nitrogen assays

Cultures were grown as described for thiosulphate assays but with 20 mM acetate and 10 mM nitrogen supplied as urea, nitrate, nitrite or ammonium (see Supplementary Fig. S5).

Acylhomoserine lactone analysis

Dichloromethane extracts of *S. pomeroyi* cultures fractionated by reverse phase thin-layer chromatography were analysed using an *Agrobacterium tumefaciens* bioassay²⁹. Expression of the LuxI homologues designated SilI1 and SilI2 was engineered by fusing a PCR-amplified complete coding sequence in line with a P_{Lac} promoter in pCR2.1-TOPO (Invitrogen). The plasmids were introduced into *E. coli* XL1-Blue cells expressing the Lac repressor *lacI*^Q. Dichloromethane extracts from an *E. coli* culture expressing SilI1 contained three AHL activities with the same migration as those identified from *S. pomeroyi* cultures, and synthesis was elevated several fold by IPTG. *E. coli* expressing SilI2 had multiple IPTG-inducible AHL activities that migrated to positions different from those activities observed in extracts of *S. pomeroyi* and from *E. coli* expressing SilI1, but still in the same general area of the TLC plate (see Supplementary Fig. S4).

Environmental library BLAST analysis

Homologues of *S. pomeroyi* genes were identified in the Sargasso Sea shotgun library¹⁰ by BLAST analysis of trace files downloaded from NCBI using a cutoff for *E* of 10^{−95} (*coxL*) and 10^{−30} (*soxB*, *nirS*, *recA*). *coxL* candidates were further screened to identify genes clustering with known CO dehydrogenases⁸ (see Supplementary Fig. S6). The number of *recA* homologues was divided by the number of lithoheterotrophy homologues to estimate frequency in cells, assuming one gene copy per cell. 16S rRNA genes from the Sargasso Sea library were assigned to taxonomic groups based on similarity in Smith–Waterman alignments with representative sequences. In the Monterey Bay BAC-end sequence library¹¹, BLAST searches yielded one *coxL*, one *soxB*, eight *recA* and no *nirS* homologues.

Phylogenetic tree construction

A phylogenetic tree of rRNA gene sequences from cultured and uncultured members of

major marine taxa was generated using the neighbour-joining method (positions 226–878, *E. coli* numbering) excluding positions with <50% conservation. Uncultured archaeon 'KTK 31A' (GenBank accession number AJ133625) served as the outgroup.

Received 22 June; accepted 1 November 2004; doi:10.1038/nature03170.

- Pomeroy, L. R. The ocean's food web: a changing paradigm. *Bioscience* **24**, 499–504 (1974).
- Giovannoni, S. J. & Rappé, M. in *Microbial Ecology of the Oceans* (ed. Kirchman, D.) 47–84 (John Wiley & Sons, New York, 2000).
- Rappé, M. S., Connon, S. A., Vergin, K. & Giovannoni, S. J. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* **418**, 630–633 (2002).
- Zengler, K. *et al.* Cultivating the uncultured. *Proc. Natl Acad. Sci. USA* **99**, 15681–15686 (2002).
- González, J. M. *et al.* Bacterial community structure associated with a dimethylsulfoniopropionate-producing North Atlantic algal bloom. *Appl. Environ. Microbiol.* **66**, 4237–4246 (2000).
- Suzuki, M. T., Béja, O., Taylor, L. T. & DeLong, E. F. Phylogenetic analysis of ribosomal RNA operons from uncultivated coastal marine bacterioplankton. *Environ. Microbiol.* **3**, 323–331 (2001).
- González, J. M. & Moran, M. A. Numerical dominance of a group of marine bacteria in the α -subclass of Proteobacteria in coastal seawater. *Appl. Environ. Microbiol.* **63**, 4237–4242 (1997).
- King, G. M. Molecular and culture-based analyses of aerobic carbon monoxide oxidizer diversity. *Appl. Environ. Microbiol.* **69**, 7257–7265 (2003).
- Zafriou, O. C., Andrews, S. S. & Wang, W. Concordant estimates of oceanic carbon monoxide source and sink processes in the Pacific yield a balanced global 'blue-water' CO budget. *Glob. Biogeochem. Cycles* **17**, 1015–1029 (2003).
- Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
- De Long, E., Heidelberg, J. F. Nelson, W. C. Monterey Bay Microbial Observatory BAC-end sequence BLAST site (<http://www.tigr.org/tdb/MBMO/>) (2000).
- Shanks, A. L. & Reeder, M. L. Reducing microzones and sulfide production in marine snow. *Mar. Ecol. Prog. Ser.* **96**, 43–47 (1993).
- Sorokin, D. Y. Oxidation of inorganic sulfur compounds by obligately organotrophic bacteria. *Microbiology* **72**, 641–653 (2003).
- Kolber, Z. S., Van Dover, C. L., Niederman, R. A. & Falkowski, P. G. Bacterial photosynthesis in surface waters of the open ocean. *Nature* **407**, 177–179 (2000).
- Béja, O., Spudich, E. N., Spudich, J. L., Leclerc, M. & DeLong, E. F. Proteorhodopsin phototrophy in the ocean. *Nature* **411**, 786–789 (2001).
- Chan, K. L. *et al.* Transcript levels of the eukaryotic translation initiation factor 5A gene peak at early G1 phase of the cell cycle in the dinoflagellate *Cryptocodinium cohnii*. *Appl. Environ. Microbiol.* **68**, 2278–2284 (2002).
- Wakeham, S. G., Lee, C. Y., Hedges, J. I., Hernes, J. J. & Peterson, M. L. Molecular indicators of diagenetic status in marine organic matter. *Geochim. Cosmochim. Acta* **61**, 5363–5369 (1997).
- Matrai, P. A. & Keller, M. D. Total organic sulfur and dimethylsulfoniopropionate (DMS) in marine phytoplankton: intracellular variations. *Mar. Biol.* **119**, 61–68 (1994).
- Riemann, L., Steward, G. F. & Azam, F. Dynamics of bacterial community composition and activity during a mesocosm diatom bloom. *Appl. Environ. Microbiol.* **66**, 578–587 (2000).
- González, J. M. *et al.* *Silicibacter pomeroyi* sp. nov. and *Roseovarius nubinhibens* sp. nov., dimethylsulfoniopropionate-demethylating bacteria from marine environments. *Int. J. Syst. Evol. Microbiol.* **53**, 1261–1269 (2003).
- Moran, M. A. & Zepp, R. G. Role of photoreactions in the formation of biologically labile compounds from dissolved organic matter. *Limnol. Oceanogr.* **42**, 1307–1316 (1997).
- Gram, L., Grossart, H.-P., Schlingloff, A. & Kiorboe, T. Possible quorum sensing in marine snow bacteria: production of acylated homoserine lactones by *Roseobacter* strains isolated from marine snow. *Appl. Environ. Microbiol.* **68**, 4111–4116 (2002).
- Klappenbach, J. A., Dunbar, J. & Schmidt, T. M. rRNA operon copy number reflects ecological strategies of bacteria. *Appl. Environ. Microbiol.* **66**, 1328–1333 (2000).
- Selje, N., Simon, M. & Brinkhoff, T. A newly discovered *Roseobacter* cluster in temperate and polar oceans. *Nature* **427**, 445–448 (2004).
- Azam, F. Microbial control of oceanic carbon flux: the plot thickens. *Science* **280**, 694–695 (1998).
- Button, D. K., Schut, F., Quang, P., Martin, R. & Robertson, B. R. Viability and isolation of marine bacteria by dilution culture: theory, procedures, and initial results. *Appl. Environ. Microbiol.* **59**, 881–891 (1993).
- Eisen, J. A. *et al.* The complete genome sequence of *Chlorobium tepidum* TLS, a photosynthetic, anaerobic, green-sulfur bacterium. *Proc. Natl Acad. Sci. USA* **99**, 9509–9514 (2002).
- Purvis, A. & Rambaut, A. Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data. *Comput. Appl. Biosci.* **11**, 247–251 (1995).
- Zhu, J., Chai, Y., Zhong, Z., Li, S. & Winans, S. C. *Agrobacterium* bioassay strain for ultrasensitive detection of N-acylhomoserine lactone-type quorum-sensing molecules: detection of autoinducers in *Mesorhizobium huakuii*. *Appl. Environ. Microbiol.* **69**, 6949–6953 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This project was funded by the US National Science Foundation. We acknowledge TIGR seqcore, informatics, and IT groups for support. We thank A. Cook for analysis of the taurine pathway and T. M. Davidsen, N. Zafar, L. Zhou, M. Gwinn and T. Creasy for bioinformatics support.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.A.M. (mmoran@uga.edu). The complete sequence has been submitted to the GenBank database under accession numbers s_pomeroyi_dss_3_267 CP000031 and s_pomeroyi_dss_3_267 CP000032.

In the platypus a meiotic chain of ten sex chromosomes shares genes with the bird Z and mammal X chromosomes

Frank Grützner¹, Willem Rens², Enkhjargal Tsend-Ayush^{1*}, Nisrine El-Mogharbel^{1*}, Patricia C. M. O'Brien², Russell C. Jones³, Malcolm A. Ferguson-Smith² & Jennifer A. Marshall Graves¹

¹Research School of Biological Sciences, Australian National University, GPO Box 475, Canberra, Australian Capital Territory 2601, Australia

²Centre for Veterinary Science, Department of Veterinary Medicine, University of Cambridge, Madingley Road, Cambridge CB3 0ES, UK

³Department of Biological Sciences, The University of Newcastle, Callaghan, New South Wales 2308, Australia

* These authors contributed equally to this work

Two centuries after the duck-billed platypus was discovered, monotreme chromosome systems remain deeply puzzling. Karyotypes of males¹, or of both sexes^{2–4}, were claimed to contain several unpaired chromosomes (including the X chromosome) that form a multi-chromosomal chain at meiosis. Such meiotic chains exist in plants⁵ and insects⁶ but are rare in vertebrates⁷. How the platypus chromosome system works to determine sex and produce balanced gametes has been controversial for decades^{1–4}. Here we demonstrate that platypus have five male-specific chromosomes (Y chromosomes) and five chromosomes present in one copy in males and two copies in females (X chromosomes). These ten chromosomes form a multivalent chain at male meiosis, adopting an alternating pattern to segregate into XXXXX-bearing and YYYYY-bearing sperm. Which, if any, of these sex chromosomes bears one or more sex-determining genes remains unknown. The largest X chromosome, with homology to the human X chromosome, lies at one end of the chain, and a chromosome with homology to the bird Z chromosome lies near the other end. This suggests an evolutionary link between mammal and bird sex chromosome systems, which were previously thought to have evolved independently.

Monotremes (mammalian subclass Prototheria) were the earliest offshoot of the mammalian lineage, diverging 210 million years ago from therian mammals (eutherians and marsupials)⁸. Only three monotremes are extant: the platypus (*Ornithorhynchus anatinus*) and two echidna species. Their phylogenetic position, and their mix of mammalian, reptilian and specialized morphological and physiological features, makes monotremes uniquely valuable for comparative genomics and for understanding sex chromosome evolution^{9,10}.

The correct chromosome number ($2n = 52$) in male and female platypus was established in 1975 (ref. 3). Measurements and banding of mitotic chromosomes revealed several that lacked obvious homologues in males. The largest, defined as the X chromosome because it was present in one copy in males and two in females, shares many genes with the eutherian and marsupial X chromosome¹⁰. However, the presence of one or more male-specific chromosomes, and of unpaired mitotic chromosomes in females, has long been controversial^{1,2,4}. No genes have been mapped to any unpaired chromosomes except for this X chromosome¹⁰, and no male-specific sequences have ever been identified.

At male meiosis in monotremes, several chromosomes assemble in a multivalent chain, analogous to meiotic chains in invertebrates that are the result of translocation heterozygosity. The X chromosome lies at one end, but the other elements are unknown and their numbers in platypus are variously reported as eight¹ or ten². How

the chain elements segregate at meiosis to produce balanced gametes is completely mysterious.

Sex determination in monotremes is another mystery. No platypus homologue of the therian testis-determining gene *SRY* (sex-determining region Y) can be identified (Supplementary Note 1).

Whereas mammals show XY male heterogamety, and testis determination is triggered by the Y-borne *SRY* gene, birds have ZW female heterogamety with no evidence of sex-specific *SRY*

sequences^{11,12}. The chicken Z chromosome includes the highly conserved *DMRT1* gene (*Drosophila doublesex* and *Caenorhabditis elegans mab-3*-related transcription factor 1)¹³. *DMRT1* maps to the Z, but not the W chromosome in emu¹⁴ as well as chicken¹³, making it a candidate for a dosage-sensitive sex-determining gene in birds. Recently a duplicated copy of *DMRT1* was identified on the medaka Y chromosome^{15,16}. However, *DMRT1* maps to human chromosome 9, and is autosomal in mammals, although two copies are required for testis differentiation¹⁷.

Both mammal and bird sex chromosome systems are believed to have evolved from autosomal pairs. An earlier proposition that the mammal XY and the bird ZW systems evolved from the same pair¹⁸ was refuted by the demonstration of homology between the chicken Z chromosome and human chromosome 9. This implied that the two systems evolved independently from different pairs of autosomes¹³.

To resolve the longstanding controversy over platypus mitotic and meiotic chromosomes we generated whole chromosome paints by flow sorting chromosomes from cultured male platypus fibroblasts, and identified them by fluorescence *in situ* hybridization (FISH, see Methods) to metaphase chromosomes of males and females¹⁹.

Eight chromosome paints (Fig. 1; Supplementary Table 1) showed different hybridization patterns in females and males, and

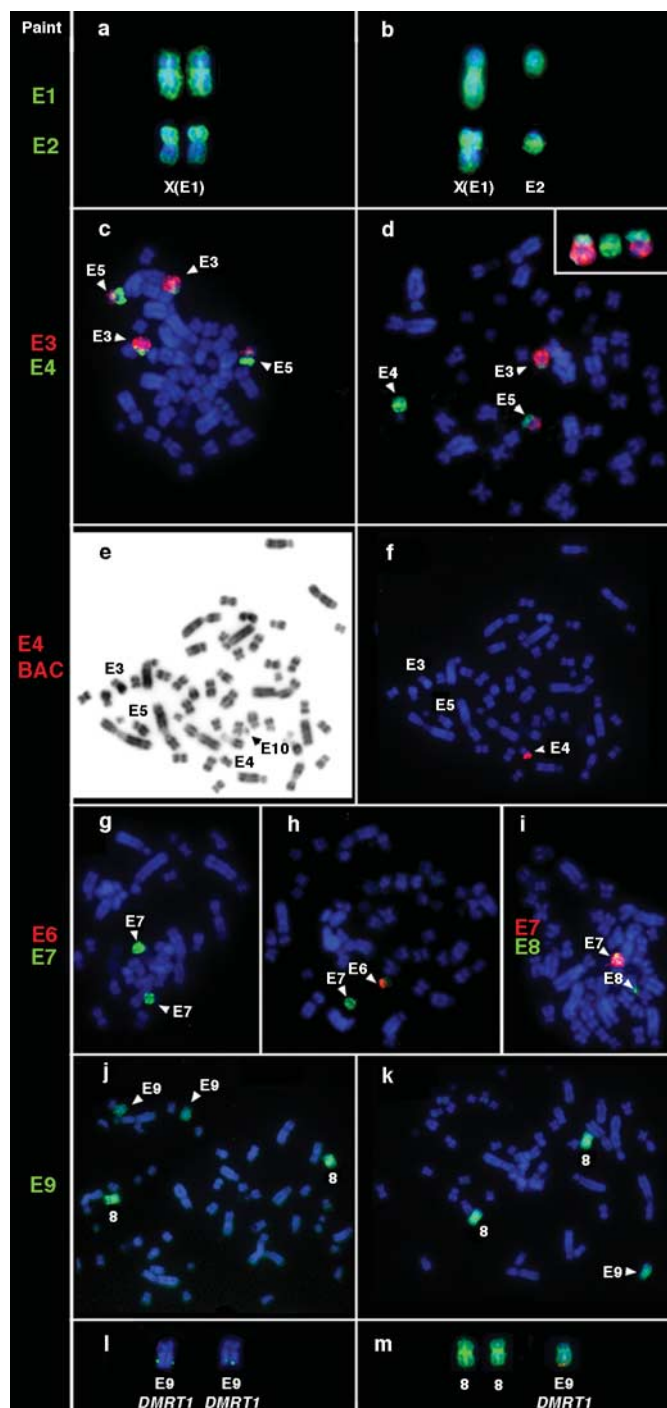


Figure 1 Fluorescence *in situ* hybridization (FISH) of chromosome paints on mitotic chromosomes. The green (FITC-labelled) and red (Cy3-labelled) E symbols indicate which paint was hybridized. **a, c, g, j, l**, FISH on female metaphase spreads; **b, d, f, h, i, k, m**, FISH on male metaphase spread. **e**, Inverted DAPI picture of the male metaphase in **f**. **l**, Localization of the *DMRT1*-containing BAC (green) in female. **m**, Co-localization of the *DMRT1*-BAC (red) and E9 (paint E9, green) in male.

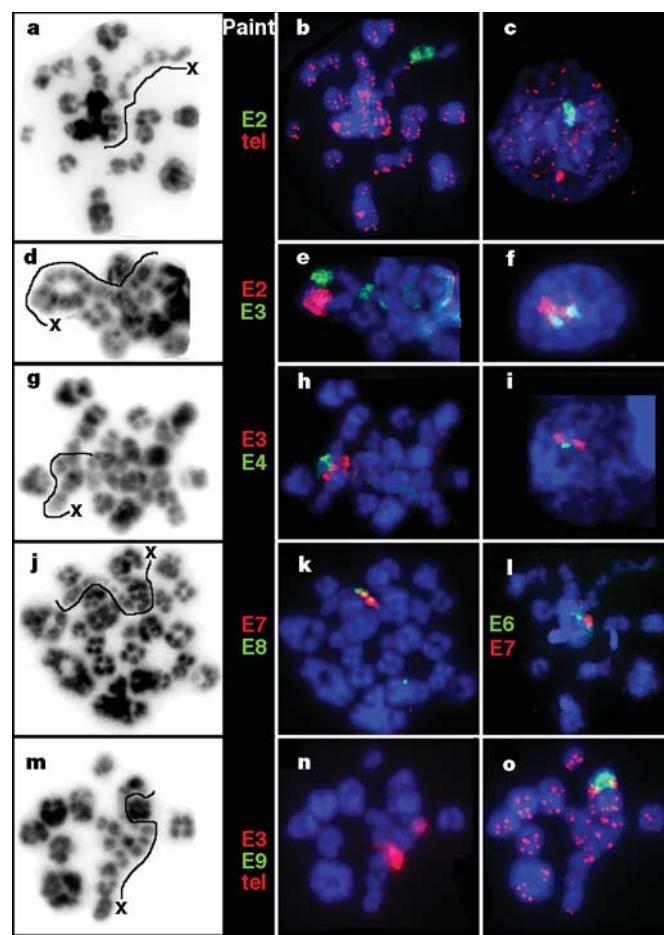


Figure 2 Order of chromosomes in the meiotic chain. Left column (**a, d, g, j, m**) shows the inverted DAPI picture of the meiotic metaphase I cells in the middle column (**b, e, h, k, n**). The chain is outlined by a black line. The start is indicated with an X. The green (FITC) and red (Cy3) labels indicate which paint was hybridized. **c**, Pachytene cell; Xp and E2 are paired. **f**, Pachytene cell; proximity of X/E2 with the unpaired E3 and E5. **i**, Pachytene cell. **l**, Metaphase I cell. **o**, Metaphase I cell; FISH with E9 and a telomeric repeat (tel).

hybridized to the meiotic chain (Fig. 2). To avoid pre-judging the identity of the chromosomes in the chain, we designated them elements E1 (the X), E2, E3 and so on. Five of these paints detected homology with other chain elements; for instance, the E2-derived paint hybridized also to E1p. The painting patterns at mitosis (Fig. 1), and their positions at meiosis (Fig. 2) identified nine unpaired chromosomes (E1–E9) in males that form the meiotic translocation chain. A tiny tenth element E10, present as the smallest chromosome in the male karyotype but absent from females, is clearly attached to E9 at the end of the chain (Fig. 1e, Fig. 3a, b and Supplementary Fig. 2). Thus males have a single copy of five X chromosomes that are paired in females, and five male-specific Y chromosomes. Females contain no unpaired mitotic chromosomes. Technically, therefore, the platypus has an $X_1X_1X_2X_2X_3X_3X_4X_4X_5X_5$ female: $X_1Y_1X_2Y_2X_3Y_3X_4Y_4X_5Y_5$ male sex chromosome system.

To discover whether the male-specific elements (Ys) harbour male-specific DNA sequences, despite their partial homology with X chromosomes, we used paint E4 to probe a platypus bacterial artificial chromosome (BAC) library. Several positive clones hybridized specifically to E4 but not E3 or E5 in males, and produced no signal at all in females, implying that E4 contains male-specific DNA (Fig. 1e, f).

To enumerate the elements of the chain, telomere repeats were hybridized to male meioses to detect chromosome ends. Chains clearly contained ten elements (Fig. 3a, b).

To determine the order of elements in the chain, paints E1–E9 were hybridized onto male meiotic metaphase I cells. This confirmed that the previously identified X chromosome lies at one end of the chain (Fig. 2a, b), its short arm paired with the male-specific E2 (Fig. 2c). Chromosome E3 lay close to E2 (Fig. 2f). These elements show no detectable homology, so they must pair over a very restricted region or form some other kind of association such as that between the X and autosomal meiotic chains in artificially constructed translocation heterozygous mice^{20,21}. Cross-hybridization between paint E3 and element E5q demonstrated homology between these elements, but chromosome order in the middle of the chain was clearly E3–E4–E5 (Fig. 2e, g–i).

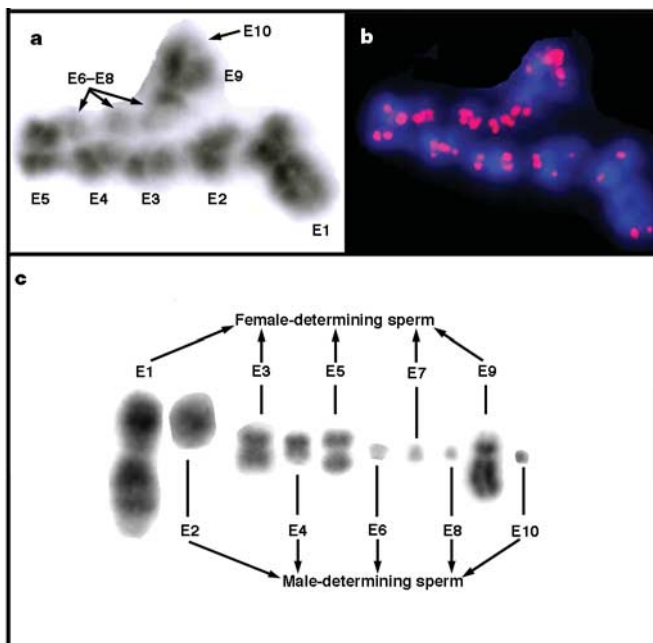


Figure 3 The elements of the chain. **a**, Inverted DAPI picture of a meiotic chain in male platypus. **b**, The same meiotic chain as in **a**, hybridized with a telomeric repeat in red (TAACCCx7). **c**, Order and segregation of the members of the chain (mitotic chromosomes).

E5 is followed in the chain by the male-specific E6, which shares some homology with the next element E7 (Figs 1h and 2l). E7 is followed by the small male-specific E8 (Fig. 2k). The chain ends with the large unpaired element E9 (Fig. 2o) to which is attached the tiny male-specific terminal element E10 (Figs 2o and 3a, b). We

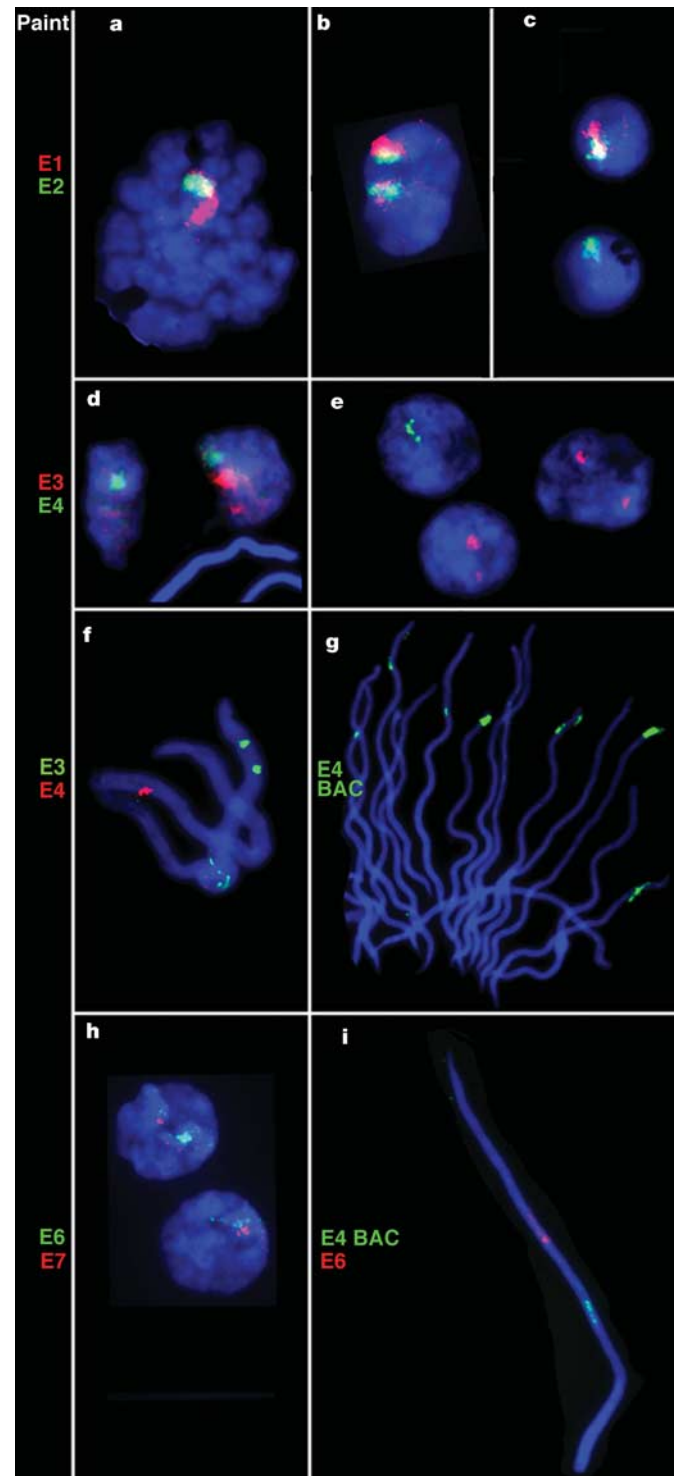


Figure 4 Alternate segregation of the chain. The column on the left indicates which paints were used, in green (FITC) or red (Cy3). **a**, Metaphase I cell. E1 and E2 are paired. **b**, Telophase I cell showing segregation of E1 and E2. **c**, Round spermatids. **d**, Meiotic telophase I cell showing segregation of E3/E5 and E4. **e**, Round spermatids. E3/E5 or E4 are present. **f**, Elongated sperm with E3/E5, or E4. **g**, Bundle of sperm showing an alternating pattern of male- and female-determining (labelled and unlabelled, respectively) sperm. **h**, Round spermatid with either E6 or E7. **i**, Elongated sperm with E4 and E6.

conclude that adjacent elements share homology, and the chain represents a complex system of sex-linked translocation heterozygosity.

Segregation of partially homologous adjacent elements would lead to unbalanced sperm. It was therefore proposed that the elements of a translocation chain undergo alternate segregation to form only two different types of spermatids containing normal or translocated elements^{1,9}. The alternating pattern of E-odd (E1, 3, 5, 7, 9) and E-even (E2, 4, 6, 8, 10) elements that we observed in the platypus chain, equivalent to alternating X and Y chromosomes, is consistent with this proposal. To demonstrate alternate segregation directly, we co-hybridized several combinations of E-even and E-odd paints onto telophase 1 cells and spermatids. E-odd and E-even paint(s) never co-located (Fig. 4 and Supplementary Table 2), implying that highly efficient alternating segregation occurs in platypus.

Fertilization of eggs (all E-odd: $X_1X_2X_3X_4X_5$) by E-odd sperm would result in an $X_1X_1X_2X_2X_3X_3X_4X_4X_5X_5$ female zygote homozygous for all X chromosomes. Fertilization by E-even ($Y_1Y_2Y_3Y_4Y_5$) sperm results in $X_1Y_1X_2Y_2X_3Y_3X_4Y_4X_5Y_5$ male zygotes with five X and five Y chromosomes (Fig. 3c).

Multiple sex chromosome systems deriving from X-autosome or Y-autosome translocations usually impose sterility in male carriers. Heterozygosity for sex chromosome-autosome translocations in mouse and humans leads to meiotic arrest or predominantly aneuploid sperm^{20,21}, although fertile $X_1Y_1X_2Y_2$ sex chromosome systems occur in a few mammalian taxa^{7,22,23}. The efficient 5X:5Y alternate segregation in platypus is therefore remarkable.

The evolution of the 5X and 5Y chromosomes of the platypus translocation chain is of considerable general interest. As for translocation complexes in invertebrates, this chain is likely to have evolved step by step, starting from a translocation between a heteromorphic sex pair and an autosome. Additional chromosome pairs became involved subsequently by serial sex chromosome-autosome translocations.

The original sex pair is therefore likely to be represented by an X and Y chromosome at one or the other end of the chain. Which end is ancestral can be assessed from the degree of homology between terminal X and Y elements. X_1 (with homology to the human X) and Y_1 , at one end of the chain, are completely homologous over the whole short arm of the X chromosome (Fig. 1a, b). This large pseudoautosomal region suggests that Y_1 is at an early stage of Y chromosome differentiation. In contrast, the tiny Y_5 at the other end of the chain is almost entirely degenerated, and its homology to X_5 is too small to be detected by our methods (Fig. 1j, k). This implies that Y_5 has undergone almost complete degeneration, and is near the endpoint proposed for heteromorphic sex chromosome systems^{10,24}. Indeed, this tiny male-specific element appears to be absent from the mitotic karyotype of the related echidna, and from its 9-membered chain²⁵. Thus we propose that the chain was initiated from translocation of an original X_5Y_5 pair with an autosome.

In the absence of any significant autosomal mapping, the identity of the original X_5Y_5 pair cannot be immediately ascertained. Speculating that the platypus sex chromosome system might retain some features of the bird system, we used *DMRT1* as a marker to determine whether X_5 and/or Y_5 has homology to the bird Z chromosome. We therefore cloned the presumed *DMRT1* gene from a platypus BAC library and sequenced two exons and three conserved intron sequences (N.E.-M. *et al.*, manuscript in preparation; Supplementary Fig. 3). We mapped *DMRT1* to male and female platypus chromosomes, and detected unambiguous signals on X_5 (Fig. 1l, m). No signal was present on Y_5 . This implies that X_5 has at least some homology to the bird Z chromosome.

We therefore conclude that the platypus sex chromosome chain was initiated from an original sex chromosome pair with homology to the bird ZW system. During early mammalian evolution, sequential translocation recruited four other autosome pairs,

most recently the pair representing the mammalian XY system.

Our observation that the two ends of the chain share homology to the mammal XY and the bird ZW systems provides the first link between these sex chromosome systems and challenges the accepted view that mammal and bird sex chromosomes evolved independently.

In the absence of *SRY*, it is unclear how monotreme sex is determined. The degenerate Y_5 is unlikely to bear a male-dominant gene because it is small and heterochromatic, and has been lost in the related echidna. E_2 at the other end is also unlikely because of its complete homology and meiotic pairing with X_p . One or more male-determining genes could be present on any of the other male-specific (Y) chromosomes. The *DMRT1* gene on chromosome X_5 , being present in two doses in females and one in males (that is, the opposite situation from birds) is unlikely to play a similar role. However, the unprecedented localization of *DMRT1* on the sex chromosomes in a mammal raises the possibility that a *DMRT1*-based sex-determining system was ancestral to all mammals. *SRY* could either have been lost in monotremes, or have evolved on the proto-Y chromosome of therian mammals after their divergence from monotremes 210 million years ago. □

Methods

Isolation of chromosome-specific painting probes

Cell culture, flow sorting, chromosome paint production and FISH were performed according to the protocol described previously²⁶. In short, chromosomes were stained with $40 \mu\text{g ml}^{-1}$ chromomycin A3 (Sigma), 2 mmol l^{-1} MgSO_4 and $2 \mu\text{g ml}^{-1}$ of Hoechst 33258 (Sigma) and incubated for at least 2 h. Ten minutes before flow analysis, sodium sulphate and sodium citrate were added to a final concentration of 10 and 25 mmol l^{-1} , respectively. The stained chromosomes were sorted on a FACStar Plus flow sorter (Becton Dickinson). Five-hundred of each chromosome type were sorted directly into separate 500- μl polymerase chain reaction (PCR) tubes containing 30 μl sterile distilled water. Flow-sorted chromosomes were used as templates for amplification or labelling by degenerate oligonucleotide-primed PCR (DOP-PCR²⁷). In the labelling reaction 50 μM dATP, dGTP and dCTP; 20 μM dTTP; and 50 μM of either biotin-16-dUTP or digoxigenin-11-dUTP were added. Chromosome paints were assigned to chromosomes by fluorescence microscopy (Leica DMRXA).

Preparation of chromosomes, meiotic cells and sperm

Mitotic metaphase chromosomes were prepared from exponentially growing platypus fibroblast cell lines from different individuals. Primary cultures were set up from toe web from previously trapped animals²⁸ as well as from four male and one female cell lines derived from animals captured at the upper Barnard river, New South Wales, Australia during breeding season (AEEC permit R.CG.07.03 (E.G.), Environment ACT permit LI 2002 270 (J.A.M.G.), NPWS permit A193 (R.C.J.)). The captured animals were killed with an intraperitoneal injection of pentobarbitone sodium (Nembutal) at a dose of 0.1 mg g^{-1} body weight. Meiotic cells and sperm were obtained by crushing the testis. The material was either directly fixed in methanol/acetic acid (3:1) or incubated in 0.075 M KCl at 37 °C as hypotonic treatment and then fixed.

Chromosome painting

Chromosome paints were hybridized to chromosomes of different individuals to exclude cell culture artefacts. For FISH, the slides were treated with $100 \mu\text{g ml}^{-1}$ RNase A/2 $\times$ SSC at 37 °C for 30 min and with 0.01% pepsin in 10 mM HCl at 37 °C for 10 min. After re-fixing for 10 min in 1 $\times$ PBS, 50 mM MgCl_2 , 1% formaldehyde, the preparations were dehydrated in an ethanol series. Slides were denatured for 2.5 min at 75 °C in 70% formamide, 2 $\times$ SSC, pH 7.0, and again dehydrated. For hybridization of one half-slide, 10 μl of biotinylated and/or digoxigenated probe DNA was coprecipitated with 10–20 μg of boiled genomic platypus DNA (as competitor), and 50 μg salmon sperm DNA (as carrier), and redissolved in 50% formamide, 10% dextran sulphate, 2 $\times$ SSC. The hybridization mixture was denatured for 10 min at 80 °C. Preannealing of repetitive DNA sequences was carried out for 30 min at 37 °C. The slides were hybridized overnight in a moist chamber at 37 °C. The slides were then washed three times for 5 min in 50% formamide, 2 $\times$ SSC at 42 °C and once for 5 min in 0.1 $\times$ SSC, pH 7.0 at 60 °C and blocked with 4 $\times$ SSC, 3% BSA, 0.1% Tween 20 at 37 °C for 30 min. Probes were detected with fluorescein isothiocyanate (FITC)-conjugated avidin and Cy3-conjugated anti-digoxigenin antibody. Chromosomes and cell nuclei were counterstained with $1 \mu\text{g ml}^{-1}$ 4,6-diamidino-2-phenylindole (DAPI) in 2 $\times$ SSC for 1 min and mounted in 90% glycerol, 0.1 M Tris-HCl, pH 8.0 and 2.3% DABCO. Images were taken with a Zeiss Axioplan epifluorescence microscope equipped with a CCD (charge-coupled device) camera (RT-Spot, Jackson Instruments), which was controlled by an Apple Macintosh computer. IPlab imaging software was used to capture grey scale images and to superimpose the source images into a colour image.

Isolation of platypus specific BAC clones and platypus *DMRT1*

A commercially available platypus BAC library (OA_Bb, Clemson University Genomics Institute, USA) was used to screen for male-specific BAC clones as well as for the platypus

DMRT1 gene. As a probe, 10 µl of DOP-PCR amplification of paint E4 were used. For *DMRT1* a 287-base-pair (bp) *Pst*I digest of chicken *DMRT1* complementary DNA, containing the DM (Doublesex and MAB-3) domain was hybridized. For both experiments the probe DNA was labelled radioactively with [³²P]-dATP and the BAC library screened according to the manufacturer's instructions (http://www.genome.clemson.edu/protocols/hyb_filter.html). Positive clones were ordered from the Clemson University Genomics Institute (CUGI). For *DMRT1*, a final positive clone was verified by PCR using primers from conserved *DMRT1* specific regions²⁹. This clone was then sequenced after subcloning into a TOPO Shotgun subcloning kit (Invitrogen) according to the manufacturer's instructions. Sequences were deposited at EMBL. The *DMRT1* BAC as well as the E4 specific BAC clones were labelled by standard nick translation with biotin or digoxigenin and 400–600 ng of labelled DNA were used to map the clones by FISH, as described above.

Received 17 June; accepted 8 September 2004; doi:10.1038/nature03021.
Published online 24 October 2004.

1. Bick, Y. in *The Meiotic Chain of Chromosomes of Monotremata* (ed. Augee, M.) 64–68 (Royal Zoological Society of New South Wales, Sydney, 1992).
2. Murtagh, C. A unique cytogenetics system in monotremes. *Chromosoma* **65**, 37–57 (1977).
3. Bick, Y. & Sharman, G. The chromosomes of the platypus (*Ornithorhynchus*): Monotremata. *Cytobios* **14**, 17–28 (1975).
4. Wrigley, J. M. & Graves, J. A. M. Karyotypic conservation in the mammalian order Monotremata (subclass Prototheria). *Chromosoma* **96**, 231–247 (1988).
5. Cleland, R. *The Cytogenetics of Oenothera* (Academic, New York, 1962).
6. Syren, R. M. & Luykx, P. Permanent segmental interchange complex in the termite *Incisitermes schwarzi*. *Nature* **266**, 167–168 (1977).
7. Fredga, K. Unusual sex chromosome inheritance in mammals. *Phil. Trans R. Soc. Lond. B* **259**, 15–36 (1970).
8. Woodburne, M. O., Rich, T. H. & Springer, M. S. The evolution of tribospheny and the antiquity of mammalian clades. *Mol. Phylogenet. Evol.* **28**, 360–385 (2003).
9. Grützner, F., Deakin, J., Rens, W., El-Mogharbel, N. & Marshall Graves, J. A. M. The monotreme genome: a patchwork of reptile, mammal and unique features? *Comp. Biochem. Physiol. A Mol. Integr. Physiol.* **136**, 867–881 (2003).
10. Graves, J. A. M. The origin and function of the mammalian Y chromosome and Y-borne genes—an evolving understanding. *Bioessays* **17**, 311–320 (1995).
11. Griffiths, R. The isolation of conserved DNA sequences related to the human sex-determining region Y gene from the lesser black-backed gull (*Larus fuscus*). *Proc. R. Soc. Lond. B* **244**, 123–128 (1991).
12. Mizuno, S. *et al.* Z and W chromosomes of chickens: studies on their gene functions in sex determination and sex differentiation. *Cytogenet. Genome Res.* **99**, 236–244 (2002).
13. Nanda, I. *et al.* 300 million years of conserved synteny between chicken Z and human chromosome 9. *Nature Genet.* **21**, 258–259 (1999).
14. Shetty, S., Kirby, P., Zarkower, D. & Graves, J. A. *DMRT1* in a ratite bird: evidence for a role in sex determination and discovery of a putative regulatory element. *Cytogenet. Genome Res.* **99**, 245–251 (2002).
15. Matsuda, M. *et al.* DMY is a Y-specific DM-domain gene required for male development in the medaka fish. *Nature* **417**, 559–563 (2002).
16. Nanda, I. *et al.* A duplicated copy of *DMRT1* in the sex-determining region of the Y chromosome of the medaka, *Oryzias latipes*. *Proc. Natl Acad. Sci. USA* **99**, 11778–11783 (2002).
17. Raymond, C. S., Murphy, M. W., O'Sullivan, M. G., Bardwell, V. J. & Zarkower, D. *Dmrt1*, a gene related to worm and fly sexual regulators, is required for mammalian testis differentiation. *Genes Dev.* **14**, 2587–2595 (2000).
18. Ohno, S. *Chromosomes and Sex Linked Genes* (Springer, Berlin, 1967).
19. Rens, W. *et al.* Resolution and evolution of the duck-billed platypus karyotype with a X₁Y₁X₂Y₂X₃Y₃X₄Y₄X₅Y₅ male sex chromosome constitution. *Proc. Natl Acad. Sci. USA* (in the press).
20. Johannisson, R. & Winking, H. Synaptonemal complexes of chains and rings in mice heterozygous for multiple Robertsonian translocations. *Chromosome Res.* **2**, 137–145 (1994).
21. Eaker, S., Pyle, A., Cobb, J. & Handel, M. A. Evidence for meiotic spindle checkpoint from analysis of spermatocytes from Robertsonian-chromosome heterozygous mice. *J. Cell Sci.* **114**, 2953–2965 (2001).
22. Toder, R. *et al.* Comparative chromosome painting between two marsupials: origins of an XX/X₁Y₁Y₂ sex chromosome system. *Mamm. Genome* **8**, 418–422 (1997).
23. Rahn, M. I., Mudry, M., Merani, M. S. & Solari, A. J. Meiotic behavior of the X₁X₂Y₁Y₂ quadrivalent of the primate *Alouatta caraya*. *Chromosome Res.* **4**, 350–356 (1996).
24. Charlesworth, B. The evolution of sex chromosomes. *Science* **251**, 1030–1033 (1991).
25. Watson, J. M., Meyne, J. & Graves, J. A. M. in *Platypus and Echidnas* (ed. Augee, M.) 53–63 (Royal Zoological Society of New South Wales, Sydney, 1992).
26. Rens, W., O'Brien, P. C., Yang, F., Graves, J. A. & Ferguson-Smith, M. A. Karyotype relationships between four distantly related marsupials revealed by reciprocal chromosome painting. *Chromosome Res.* **7**, 461–474 (1999).
27. Telenius, H. *et al.* Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* **13**, 718–725 (1992).
28. Wrigley, J. M. & Graves, J. A. M. Two monotreme cell lines, derived from female platypuses (*Ornithorhynchus anatinus*; Monotremata, Mammalia). *In Vitro* **20**, 321–328 (1984).
29. Brunner, B. *et al.* Genomic organization and expression of the doublesex-related gene cluster in vertebrates and detection of putative regulatory regions for *DMRT1*. *Genomics* **77**, 8–17 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank P. Miethke and D. McMillan for technical assistance, D. Zarkower for providing the chicken *DMRT1* cDNA and A. Pask for information on *SOX* genes and *SRY* in monotremes, and provision of Supplementary Fig. 1. This work was supported by the Australian Research Council (E.G. and J.A.M.G.). W.R. and M.A.F.-S. are supported by a Wellcome Trust grant to the Cambridge Resource Centre for Comparative Genomics. Facilities were provided by Macquarie Generation and Glenrock Station, NSW. Approval to collect animals was granted by the New South Wales National Parks and Wildlife Services, New South Wales Fisheries,

Environment Australian Capital Territory and the Animal Experimentation Ethics Committee, Australian National University.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.G. (frank.gruetzner@anu.edu.au). The sequences reported in this paper have been deposited in EMBL under accession numbers AJ744847, AJ744848, AJ744849.

Lipocalin 2 mediates an innate immune response to bacterial infection by sequestering iron

Trude H. Flo^{1,2*}, Kelly D. Smith^{1,3*}, Shintaro Sato⁴, David J. Rodriguez¹, Margaret A. Holmes⁵, Roland K. Strong⁵, Shizuo Akira⁴ & Alan Aderem¹

¹Institute for Systems Biology, Seattle, Washington 98103, USA

²Institute for Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, 7489 Trondheim, Norway

³University of Washington, Department of Pathology, Seattle, Washington 98195, USA

⁴Department of Host Defense, Research Institute for Microbial Diseases, Osaka University, and Exploratory Research for Advanced Technology (ERATO), Japan Science and Technology Corporation, Osaka 565-0871, Japan

⁵Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA

* These authors contributed equally to this work

Although iron is required to sustain life, its free concentration and metabolism have to be tightly regulated¹. This is achieved through a variety of iron-binding proteins including transferrin and ferritin². During infection, bacteria acquire much of their iron from the host by synthesizing siderophores that scavenge iron and transport it into the pathogen^{3,4}. We recently demonstrated that enterochelin, a bacterial catecholate siderophore,

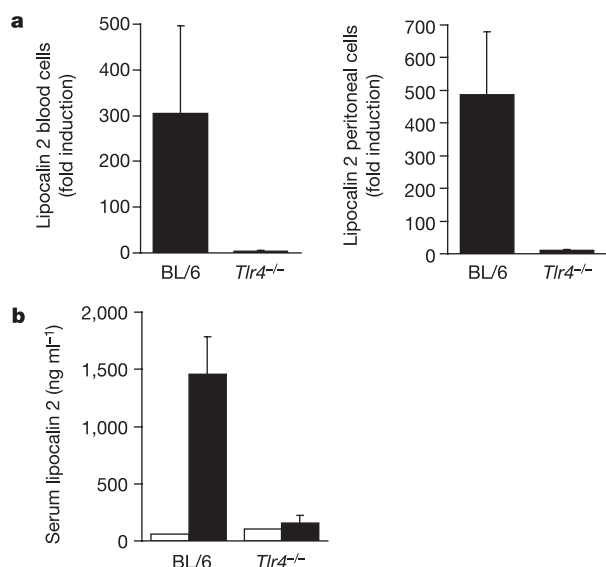


Figure 1 Lipocalin 2 production is induced through TLRs. Lipocalin 2 mRNA and protein measured 4 h after i.p. injection of C57BL/6 wild-type and TLR4-deficient mice with 10 µg LPS (*n* = 3 mice). **a**, Lipocalin 2 mRNA from blood cells and peritoneal cells is shown as fold induction values relative to the mRNA from PBS-injected mice. **b**, Serum levels of lipocalin 2 protein after LPS injection (filled bars) are compared to the levels in PBS-injected mice (open bars). Errors bars show the s.d. for each experiment.

binds to the host protein lipocalin 2 (ref. 5). Here, we show that this event is pivotal in the innate immune response to bacterial infection. Upon encountering invading bacteria the Toll-like receptors on immune cells stimulate the transcription, translation and secretion of lipocalin 2; secreted lipocalin 2 then limits bacterial growth by sequestering the iron-laden siderophore. Our finding represents a new component of the innate immune system and the acute phase response to infection.

In mammals, distinct combinations of at least ten Toll-like receptors (TLRs) discriminate between the large number of microbial components found in nature⁶. Activation of the mammalian TLRs leads to the induction of inflammatory responses and to the development of adaptive immunity⁷. Complementary DNA microarray analysis demonstrates that lipocalin 2 (also known as neutrophil gelatinase-associated lipocalin, siderocalin, 24p3, or uterocalin^{5,8-12}) transcription is markedly increased in macrophages stimulated *in vitro* with lipopolysaccharide (LPS; detected by TLR4 (ref. 13)), Pam₃CSK₄ lipopeptide (TLR2/1 (refs 14–16)) and flagellin (TLR5 (ref. 17)) (data not shown). *In vivo*, LPS induces a

200-fold increase in lipocalin 2 messenger RNA transcription and a 20-fold increase in protein concentration in a TLR4-dependent manner (Fig. 1a, b).

To assess the role of lipocalin 2 *in vivo*, we generated lipocalin-2-deficient mice (Supplementary Fig. 1). These mice have normal litters and no apparent phenotype when housed in specific pathogen-free conditions. However, intraperitoneal challenge with a sub-lethal dose of a clinical strain of *Escherichia coli*, H9049, results in substantial increases in bacteraemia and bacterial burden in the liver and spleen (Fig. 2a). The peak bacteraemia was on average 1,000-fold ($P = 0.02$) greater in lipocalin-2-deficient mice than in controls, and significant differences in bacterial counts in the liver ($P = 0.04$) and spleen ($P = 0.03$) were also seen (Fig. 2a and not shown). The physical signs of sepsis (lethargy, hunched posture, ruffled fur) correlated well with the bacterial burden, especially in the blood. Despite the marked difference in bacteraemia between wild-type and lipocalin-2-deficient mice there were no apparent differences in leukocyte numbers (Supplementary Table 1) or in neutrophil infiltration and tissue iron distribution (Supplementary

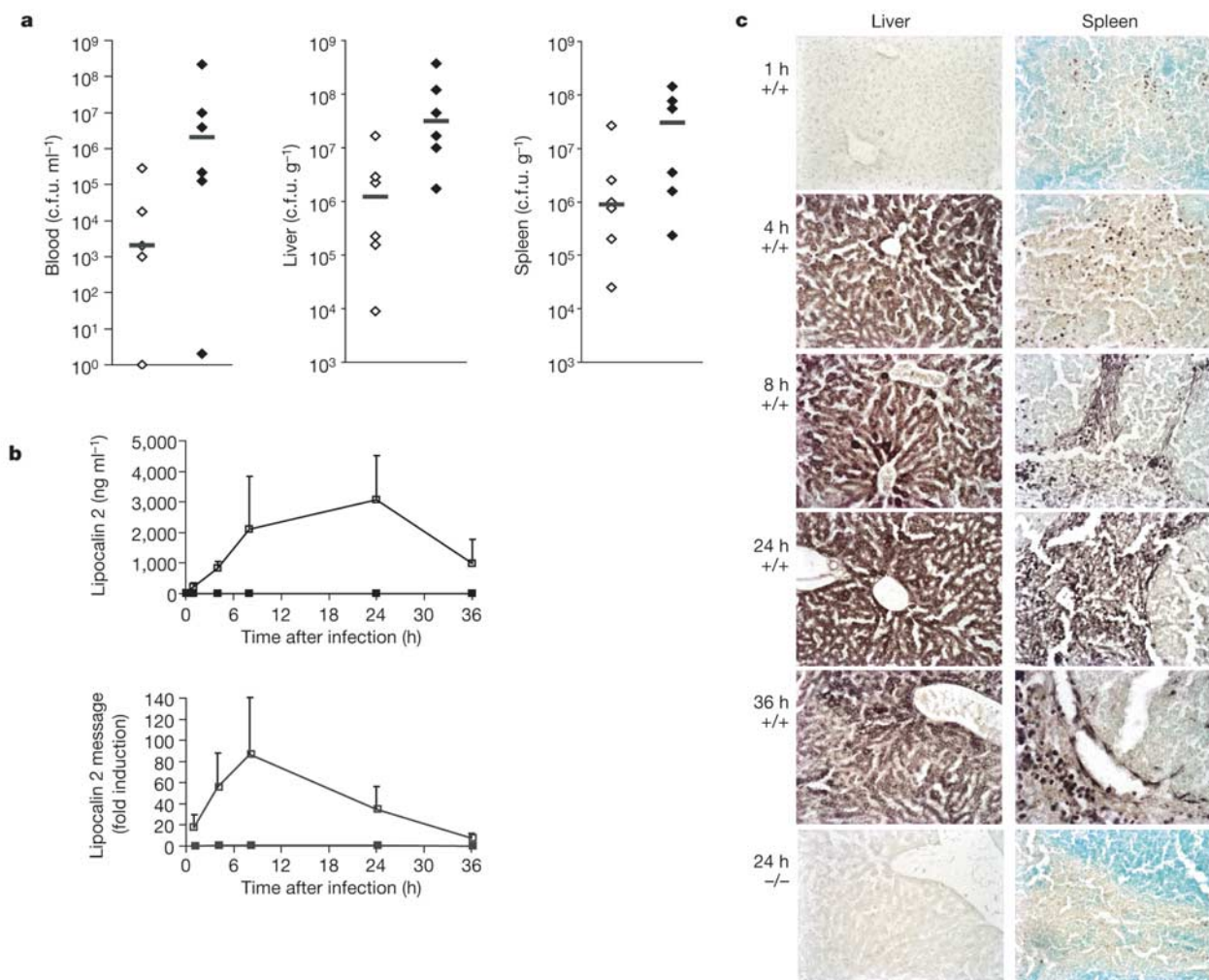
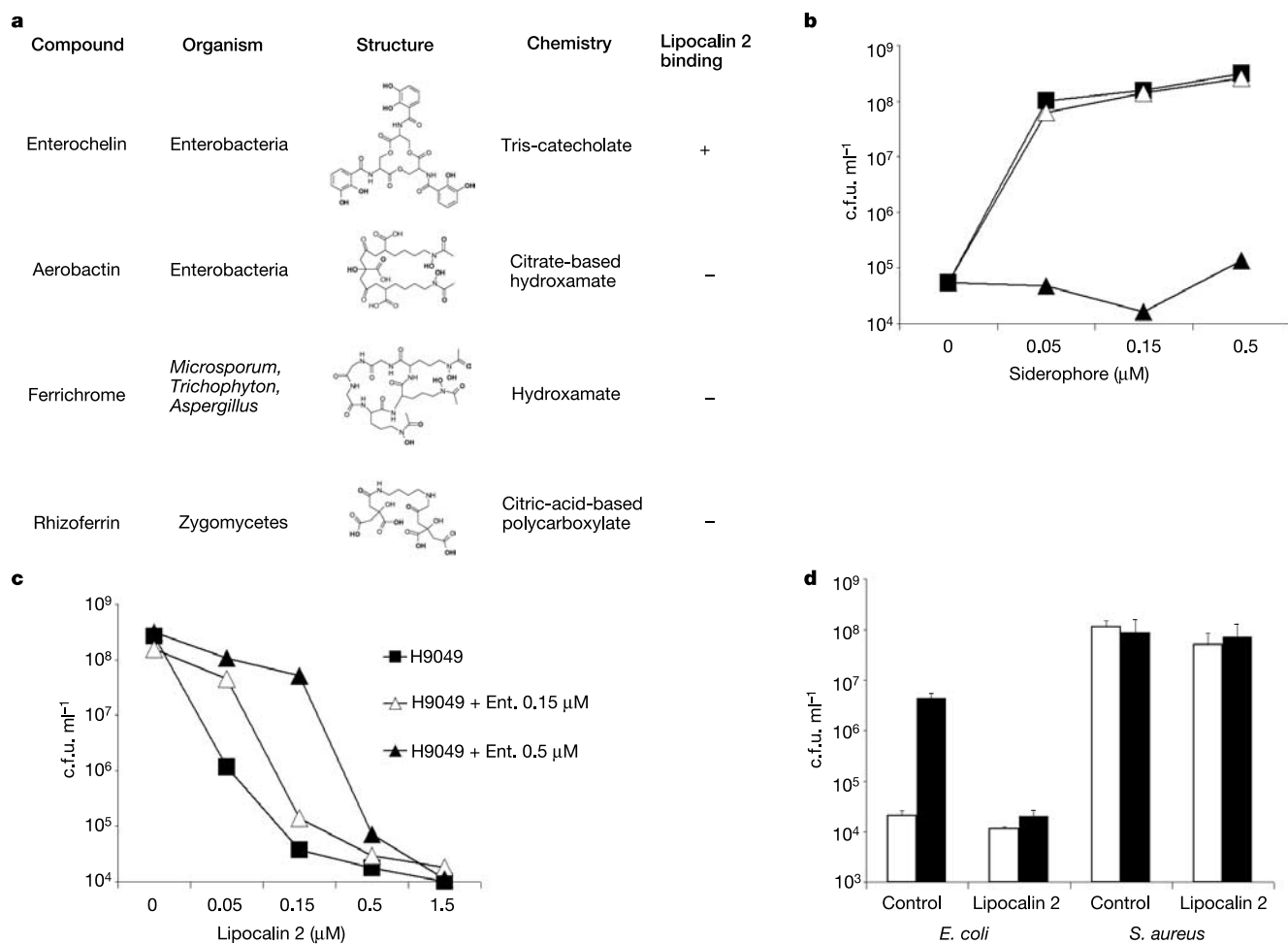


Figure 2 Lipocalin 2 protects against infection with *E. coli*. **a**, Bacterial burden in blood, liver and spleen 36 h after i.p. infection of wild type (open diamonds) and lipocalin-2-deficient (filled diamonds) mice with 0.6×10^8 c.f.u. *E. coli* H9049. Individual c.f.u. values and the median (horizontal line) from one representative experiment out of three are shown ($n = 6$ mice for each group, $P < 0.05$ in the three experiments combined, as evaluated by two-sided Student's *t*-test and Fisher's method for combining *P*-values). **b**, **c**, Induction of lipocalin 2 in serum, blood cells and tissues 1, 4, 8, 24 and 36 h after i.p. infection of lipocalin-2-deficient and wild-type mice with 0.6×10^8 c.f.u. *E. coli* H9049

($n = 5$ mice in each group for each time point). **b**, Serum levels of lipocalin 2 protein (top panel) and lipocalin 2 mRNA in blood cells (bottom panel) measured in wild-type mice (open squares) and in lipocalin-2-deficient mice (closed squares). Data are presented as the mean and s.d. for five mice at each time point. **c**, Tissue induction of lipocalin 2 in wild-type mice detected on frozen sections of liver and spleen with a polyclonal antibody specific for lipocalin 2. From the lipocalin 2 knockouts, only the 24-h time point is shown (bottom panel).

The basal serum concentration of lipocalin 2 in uninfected, control mice was approximately 100 ng ml⁻¹ (Fig. 2b, top panel).

Serum represents an iron-restricted environment, as most of the iron is complexed with transferrin. Invading pathogens respond to



a, Binding of lipocalin 2 to siderophores from different sources and chemical classes. Consistent with the pattern and high degree of sequence identity between human and mouse lipocalin 2, both orthologues display a conserved binding specificity for catechol-type siderophores. **b**, Increasing concentrations of enterochelin (squares) and ferrichrome (open triangles), but not aerobactin (filled triangles), enhance the growth of *E. coli* H9049 in an iron-restricted environment (RPM/10% FBS). **c**, Recombinant

Publishing Group 919

type siderophores such as aerobactin and ferrichrome, or to polycarboxylate-type siderophores such as rhizoferrin (Fig. 3a). This indicates that lipocalin 2 is effective against bacteria that depend exclusively on catecholate-type siderophores for growth in iron-restricted environments. The H9049 strain of *E. coli* used in these studies can use enterochelin and ferrichrome, but not aerobactin, to import iron (Fig. 3b). Consistent with this finding H9049 does not possess the genes for aerobactin biosynthesis (*iucD*) or uptake (*iutA*) (data not shown). As predicted, lipocalin 2 inhibited the growth of H9049 in a dose-dependent manner (Fig. 3c). It also inhibited the growth of H9049 in the presence of exogenously added enterochelin, and this inhibition was only overcome when the enterochelin concentration exceeded that of lipocalin 2 (Fig. 3c), reflecting their equimolar stoichiometry of binding⁵. Taken together, the data demonstrate that the bacteriostatic effect of lipocalin 2 is dependent on its sequestration of enterochelin.

As the most pronounced difference in bacterial burden between infected wild-type and lipocalin-2-deficient mice was seen in the blood, we tested whether bacterial growth was limited by the increased serum lipocalin levels found in wild-type mice. H9049 cultures grew rapidly in acute phase serum from lipocalin-2-deficient mice, reaching levels 1,000-fold higher than cultures grown in acute phase serum from wild-type mice (Fig. 3d). The addition of recombinant lipocalin 2 to the acute phase serum of the knockout mice restored the bacteriostatic activity (Fig. 3d). Acute phase serum from wild-type mice contained 40 nM ($1 \mu\text{g ml}^{-1}$) of lipocalin 2, and an equivalent concentration of recombinant lipocalin 2 conferred 99% growth inhibition of the H9049 strain (Fig. 3c).

The bacteriostatic effect of lipocalin 2 was absolutely specific for bacteria that acquire iron through lipocalin-2-binding siderophores; for example, enterochelin (Fig. 3d) and mycobactin (R. Strong, unpublished data). By contrast, lipocalin 2 deficiency had no effect on bacteria that acquire iron through siderophores that do not bind lipocalin 2; for example, *Staphylococcus aureus*,

which does not require enterochelin-type siderophores (Fig. 3d).

To test the role of lipocalin 2 in an acute lethal infection, we increased the bacterial challenge (2×10^8 colony-forming units (c.f.u.) *E. coli* H9049) and compared the survival of lipocalin-2-deficient mice and wild-type controls (Fig. 4a). Knockout mice showed substantially accelerated lethality: within the first 42 h greater than 80% of lipocalin-2-deficient mice had died compared with 0% of wild-type mice (Fig. 4a). Only 17% of the lipocalin-2-deficient mice recovered completely from the infection, whereas all of the wild-type mice survived.

To explore further the specificity of the protective effect of lipocalin 2 against enterochelin-dependent bacterial infection, we supplied the bacteria with a lipocalin-2-independent route of iron acquisition by adding ferrichrome, a hydroxamate-type siderophore. Ferrichrome is not synthesized by *E. coli*, but can be used by the bacterium for iron acquisition (Fig. 3b). We first showed *in vitro* that ferrichrome is able to circumvent lipocalin-2-dependent bacteriostatic activity (Fig. 4b). We next challenged wild-type mice with intraperitoneal infection of *E. coli* H9049, with or without co-injection of ferrichrome. Mice that received bacteria and ferrichrome demonstrated substantially increased lethality, similar to the lethality of lipocalin-2-deficient mice; ferrichrome alone had no effect (Fig. 4c). Further insight into the specificity of lipocalin 2 function was provided by the observation that lipocalin 2 deficiency has no effect on the survival of mice intraperitoneally infected with *S. aureus* (Fig. 4d), a bacterium whose acquisition of iron is lipocalin-2-independent (Fig. 3d). These results demonstrate that lipocalin 2 confers resistance to bacterial infection by abrogating enterochelin-like, siderophore-dependent iron uptake by bacteria. This host defence mechanism is important because a great many human pathogens including *E. coli*, *Salmonella* spp., *Brucella abortus*, *Bacillus anthracis*, *Burkholderia cepacia*, *Corynebacterium diphtheriae*, *Paracoccus* spp. and *Vibrio* spp. make enterochelin-like siderophores. In addition, the near absolute conservation of residues contributing to the siderophore-binding pocket of lipocalin 2

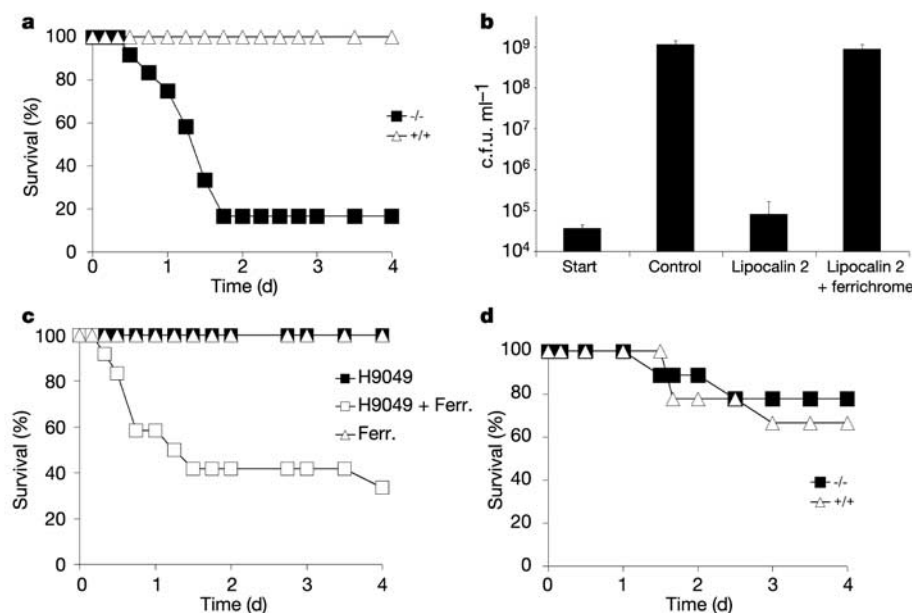


Figure 4 Lipocalin 2 protection against bacterial infection is siderophore specific. **a**, Survival curve comparing lipocalin-2-deficient mice and wild-type controls after i.p. challenge with 2×10^8 c.f.u. *E. coli* H9049 per mouse ($n = 13$ for each group). **b**, Growth of bacteria in an iron-restricted environment (RPMI/10% FBS) in the presence of added lipocalin 2 (1.5 μM) and ferrichrome (5 nM), as indicated. Errors bars show the s.d. for each experiment. **c**, Survival curve comparing wild-type mice after i.p. challenge

with 2×10^8 c.f.u. *E. coli* H9049 per mouse with or without co-injection of ferrichrome (25 nmol) ($n = 12$ for each group), or for wild-type mice ($n = 6$) injected with ferrichrome (25 nmol) alone. **d**, Survival curve comparing lipocalin-2-deficient and wild-type mice after i.p. challenge with 3×10^8 *S. aureus* ($n = 9$ for each group). No mice died after i.p. infection with 1×10^8 *S. aureus*, whereas all mice died after i.p. infection with 10^9 *S. aureus* (data not shown).

(ref. 5) suggests that this defence mechanism is conserved across species.

The expression of lipocalin 2 in multiple tissues outside the setting of infection suggests that it may have other functions, such as have been postulated in kidney development²¹ and the implantation and parturition stages of pregnancy^{12,22}. Although we cannot exclude a role for lipocalin 2 in these functions, the phenotype of lipocalin-2-deficient mice argues against an obligatory role in these biological processes. Alternatively, there may be subtle effects that will only be revealed by more detailed investigation of lipocalin-2-deficient mice.

Lipocalin 2 has also been implicated as a factor that induces apoptosis through interaction with an as yet undefined receptor²³. It is formally possible that lipocalin-2-deficient mice are more susceptible to infection because of a defect in apoptosis. However, two findings argue against this possibility. First, our experiments demonstrate that lipocalin 2 deficiency renders mice susceptible to bacteria that depend on enterochelin-based iron uptake (Fig. 4). Second, we have not detected an apoptotic defect in lipocalin-2-deficient mice (Supplementary Fig. 3).

We have described a new mechanism by which the innate immune system curbs bacteraemia during the early stage of infection. Upon encountering bacteria, innate immune cells produce and secrete lipocalin 2, which, in turn, limits bacterial growth by iron sequestration. Our data in the mouse model indicate that the absence of this defence mechanism can lead to sepsis and death; these findings, therefore, have substantial implications for the clinical treatment of bacteraemia and bacterial sepsis. □

Methods

Generation of lipocalin-2-deficient mice

The genomic DNA fragment containing the *Lcn2* gene was screened from a 129/Sv mouse genomic library and characterized by restriction enzyme mapping and sequence analysis. The targeting vector was constructed by replacing a 1.9-kilobase fragment of the *Lcn2* gene, including exons 2–5, with a neomycin-resistance gene cassette. A herpes simplex virus thymidine kinase gene, driven by an MC1 promoter, allowed for negative selection. The targeting vector was transfected into embryonic day (E)14.1 embryonic stem (ES) cells. Targeted ES cells, identified by polymerase chain reaction (PCR) and Southern blotting, were subsequently injected into C57BL/6 blastocysts. Male chimaeric mice were then mated to C57BL/6 female mice and the heterozygote F₁ progeny were intercrossed to generate lipocalin-2-deficient mice. Knockout mice and their wild-type littermates were then used in separate breeding programmes to generate populations of lipocalin 2 knockouts and wild-type controls for subsequent experiments.

Lipocalin 2 mRNA and protein measurements

Lipocalin 2 messenger RNA was measured in freshly drawn blood cells and extracted peritoneal cells using a PAXgene blood RNA kit (PreAnalytix) and an RNeasy mini kit (Qiagen), respectively, both in accordance with the manufacturer's protocols. cDNA was synthesized and quantified by real-time PCR using a lipocalin-2-specific probe and primer set. All data were normalized to a housekeeping gene (*Efla*) measured in the same sample.

For detection of lipocalin 2 protein we generated polyclonal antibodies by immunizing rabbits with recombinant mouse lipocalin 2 protein⁸ (R&R Research and Development). Affinity-purified antibody was used both as capture (4 µg ml⁻¹) and detection (biotinylated, 0.1 µg ml⁻¹) antibody in a sandwich ELISA. The detection limit was 100–150 pg ml⁻¹ lipocalin 2. The same anti-lipocalin 2 antibody was used in western blots at 1 µg ml⁻¹ and in immunohistochemistry on frozen sections at 3 µg ml⁻¹.

Mouse intraperitoneal injections and infection model

TLR4 (backcrossed eight generations onto the C57BL/6 background) or lipocalin-2-deficient mice were age- and sex-matched with C57BL/6 or lipocalin 2 wild-type controls, respectively. The mice were kept in specific pathogen-free housings. For measurements of lipocalin 2 protein and mRNA, TLR4-deficient mice and C57BL/6 controls were injected intraperitoneally (i.p.) with 10 µg LPS (List Biological Laboratories) in 0.5 ml PBS. The mice were killed 4 h later and the blood was used for RNA extraction and for generation of serum. Acute phase serum was produced by i.p. injection of 10⁸ c.f.u. heat-killed *E. coli* (clinical isolate H9049; provided by S. Swanzey) into lipocalin 2 wild-type and knockout mice, and after 5 h mice were killed and bled. For infections with live bacteria, log phase *E. coli* H9049 or *S. aureus* (American Type Culture Collection 25923) were washed, resuspended in PBS to the desired concentration and injected i.p. in 0.5 ml per mouse using a 30-gauge needle. All injected doses were verified by counting c.f.u. Mouse behaviour was carefully monitored every 12 h. During survival experiments the mice were not left alone for more than 3–4 h. All dead mice were analysed by necropsy. For measuring the bacterial burden, heparinized blood samples and homogenized liver and spleen were diluted in PBS and plated on LB agar to determine c.f.u. Blood samples were also used for RNA extraction and measurements of serum lipocalin 2 protein. Tissue samples were fixed for paraffin sectioning and immunohistochemistry (frozen sections).

Siderophore binding to lipocalin 2

Binding of siderophores to either human or mouse lipocalin 2 was determined as previously reported⁵ or by a surface plasmon resonance assay, in HBS-P buffer (10 mM HEPES, pH 7.4, 150 mM NaCl, 0.005% P-20) using a Biacore 3000 system (Biacore AB). Siderophores were obtained from Sigma-Aldrich, Biophore Research Products or as a gift of K. N. Raymond. We classified a binding interaction as an interaction with a dissociation constant (*K_d*) estimated at less than 100 nM; non-binding siderophores interact with protein with a *K_d* estimated to be greater than 1 mM. The kinetics of binding are complex and multiphasic, currently obviating quantification of the affinities. Owing to the marked solution instability of enterochelin, a non-hydrolysable synthetic analogue²⁴, differing only in backbone chemistry, was used as a surrogate in the binding assay.

In vitro growth assay

For *in vitro* growth measurements of *E. coli* H9049 or *S. aureus* we added 10³–10⁴ c.f.u. of log phase bacteria in 100 µl per well, 96-well plates. The bacteria were grown in RPMI with 10–20% heat-inactivated fetal bovine serum (FBS) or acute phase serum from lipocalin 2 wild-type or knockout mice. Purified siderophores (EMC microcollections GmbH) and/or recombinant mouse lipocalin 2 (ref. 4) were added at the indicated concentrations, and growth was measured as c.f.u. from serial dilutions plated onto LB agar.

Received 3 August; accepted 11 October 2004; doi:10.1038/nature03104.

Published online 7 November 2004.

- Weinberg, E. D. Iron withholding: a defense against infection and neoplasia. *Physiol. Rev.* **64**, 65–102 (1984).
- Andrews, N. C. Iron homeostasis: insights from genetics and animal models. *Nature Rev. Genet.* **1**, 208–217 (2000).
- Ratledge, C. & Dover, L. G. Iron metabolism in pathogenic bacteria. *Annu. Rev. Microbiol.* **54**, 881–941 (2000).
- Faraldo-Gomez, J. D. & Sansom, M. S. Acquisition of siderophores in gram-negative bacteria. *Nature Rev. Mol. Cell Biol.* **4**, 105–116 (2003).
- Goetz, D. H. *et al.* The neutrophil lipocalin NGAL is a bacteriostatic agent that interferes with siderophore-mediated iron acquisition. *Mol. Cell* **10**, 1033–1043 (2002).
- Takeda, K., Kaisho, T. & Akira, S. Toll-like receptors. *Annu. Rev. Immunol.* **21**, 335–376 (2003).
- Janeway, C. A. Jr & Medzhitov, R. Innate immune recognition. *Annu. Rev. Immunol.* **20**, 197–216 (2002).
- Kjeldsen, L., Cowland, J. B. & Borregaard, N. Human neutrophil gelatinase-associated lipocalin and homologous proteins in rat and mouse. *Biochim. Biophys. Acta* **1482**, 272–283 (2000).
- Triebl, S., Blaser, J., Reinke, H. & Tschesche, H. A 25 kDa alpha 2-microglobulin-related protein is a component of the 125 kDa form of human gelatinase. *FEBS Lett.* **314**, 386–388 (1992).
- Kjeldsen, L., Johnsen, A. H., Sengelov, H. & Borregaard, N. Isolation and primary structure of NGAL, a novel protein associated with human neutrophil gelatinase. *J. Biol. Chem.* **268**, 10425–10432 (1993).
- Flower, D. R., North, A. C. & Attwood, T. K. Mouse oncogene protein 24p3 is a member of the lipocalin protein family. *Biochem. Biophys. Res. Commun.* **180**, 69–74 (1991).
- Liu, Q., Ryon, J. & Nilsen-Hamilton, M. Uterocalin: a mouse acute phase protein expressed in the uterus around birth. *Mol. Reprod. Dev.* **46**, 507–514 (1997).
- Poltorak, A. *et al.* Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in Tlr4 gene. *Science* **282**, 2085–2088 (1998).
- Lien, E. *et al.* Toll-like receptor 2 functions as a pattern recognition receptor for diverse bacterial products. *J. Biol. Chem.* **274**, 33419–33425 (1999).
- Ozinsky, A. *et al.* The repertoire for pattern recognition of pathogens by the innate immune system is defined by cooperation between toll-like receptors. *Proc. Natl Acad. Sci. USA* **97**, 13766–13771 (2000).
- Takeuchi, O. *et al.* Cutting edge: role of Toll-like receptor 1 in mediating immune response to microbial lipoproteins. *J. Immunol.* **169**, 10–14 (2002).
- Hayashi, F. *et al.* The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. *Nature* **410**, 1099–1103 (2001).
- Xu, S. & Venge, P. Lipocalins as biochemical markers of disease. *Biochim. Biophys. Acta* **1482**, 298–307 (2000).
- Liu, Q. & Nilsen-Hamilton, M. Identification of a new acute phase protein. *J. Biol. Chem.* **270**, 22565–22570 (1995).
- Winkelmann, G. Microbial siderophore-mediated transport. *Biochem. Soc. Trans.* **30**, 691–696 (2002).
- Yang, J. *et al.* An iron delivery pathway mediated by a lipocalin. *Mol. Cell* **10**, 1045–1056 (2002).
- Suire, S., Stewart, F., Beauchamp, J. & Kennedy, M. W. Uterocalin, a lipocalin provisioning the preattachment equine conceptus: fatty acid and retinol binding properties, and structural characterization. *Biochem. J.* **356**, 369–376 (2001).
- Devireddy, L. R., Teodoro, J. G., Richard, F. A. & Green, M. R. Induction of apoptosis by a secreted lipocalin that is transcriptionally regulated by IL-3 deprivation. *Science* **293**, 829–834 (2001).
- Hou, Z., Stack, T. D. P., Sunderland, C. J. & Raymond, K. N. Enhanced iron (III) chelation through ligand predisposition: Syntheses, structures and stability of Tris-catecholate enterobactin analogs. *Inorg. Chim. Acta* **263**, 341–355 (1997).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We acknowledge T. Hawn and M. Matsumoto for discussions. This work was supported by grants from The Norwegian Research Council (to T.H.E.) and the NIH (to A.A., R.K.S. and K.D.S.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.A. (aaderem@systemsbiology.org).

Modelling how ribavirin improves interferon response rates in hepatitis C virus infection

Narendra M. Dixit¹, Jennifer E. Layden-Almer², Thomas J. Layden² & Alan S. Perelson¹

¹Theoretical Biology and Biophysics, Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

²Department of Medicine, University of Illinois at Chicago, Chicago, Illinois 60612, USA

Nearly 200 million individuals worldwide are currently infected with hepatitis C virus (HCV)¹. Combination therapy with pegylated interferon and ribavirin, the latest treatment for HCV infection, elicits long-term responses in only about 50% of patients treated^{2–4}. No effective alternative treatments exist for non-responders⁵. Consequently, significant efforts are continuing to maximize response to combination therapy^{6,7}. However, rational therapy optimization is precluded by the poor understanding of the mechanism(s) of ribavirin action against HCV⁸. Ribavirin alone induces either a transient early decline or no decrease in HCV viral load^{9–12}, but in combination with interferon it significantly improves long-term response rates^{2–4,13–15}. Here we present a model of HCV dynamics in which, on the basis of growing evidence^{16–21}, we assume that ribavirin decreases HCV infectivity in an infected individual in a dose-dependent manner. The model quantitatively predicts long-term response rates to interferon monotherapy and combination therapy, fits observed patterns of HCV RNA decline in patients undergoing therapy, reconciles conflicting observations of the influence of ribavirin on HCV RNA decline, provides key insights into the mechanism of ribavirin action against HCV, and establishes a framework for rational therapy optimization.

Ribavirin, a purine analogue, is phosphorylated within cells and incorporated into the RNA of replicating virions, thereby increasing the mutation frequency and reducing the specific infectivity of new virions²¹. We therefore assume that ribavirin (alone or in combination with interferon) renders a fraction of newly produced virions non-infectious and write the following equations to describe HCV dynamics in an individual undergoing combination therapy:

$$\frac{dI}{dt} = \beta TV_1 - \delta I \quad (1)$$

$$\frac{dV_1}{dt} = (1 - \rho)(1 - \varepsilon)pI - cV_1 \quad (2)$$

$$\frac{dV_{NI}}{dt} = \rho(1 - \varepsilon)pI - cV_{NI} \quad (3)$$

Infectious HCV virions, V_1 , infect target cells, T , to create productively infected cells, I , at rate βTV_1 . In the absence of therapy, each productively infected cell releases new virions at rate p . Interferon lowers p by a factor $(1 - \varepsilon)$, where ε is the effectiveness of interferon^{22,23}. Of the virions released, we assume that ribavirin renders a fraction ρ non-infectious, giving rise to the population V_{NI} . We call ρ the effectiveness of ribavirin. Productively infected cells are lost with first-order rate constant δ . Free virions are cleared from plasma with first-order rate constant c . The measured viral load $V = V_1 + V_{NI}$.

We present in Fig. 1 viral load decay profiles obtained by the numerical integration of equations (1)–(3) for three values of the loss rate of productively infected cells, $\delta = 0.01 \text{ d}^{-1}$, 0.14 d^{-1} and 0.4 d^{-1} , spanning the range estimated^{22–24}. To mimic the slow accumulation of ribavirin in plasma, we let ribavirin effectiveness

$\rho = \rho_{\max}(1 - \exp(-t/t_a))$, with the accumulation timescale $t_a = 5.6 \text{ d}$ (ref. 25). Ribavirin effectiveness, ρ , thus increases from 0 at $t = 0$ to the asymptotic value, $\rho_{\max}$, in $t \approx 28 \text{ d}$. Models assuming constant interferon effectiveness, ε , have successfully predicted plasma HCV RNA decline in patients under interferon therapy^{22,23}.

With constant ε , we find that V exhibits a biphasic decay with a fast first phase that lasts 1–2 d from the onset of therapy and a slow second phase for the rest of the treatment period. Interestingly, when ε is high, ribavirin has negligible influence on viral load decay. Theoretically, after two months of therapy with $\varepsilon = 0.95$, the difference in V when $\rho_{\max} = 0$ and 1 is less than $0.1 \log$ (1 log corresponds to a 10-fold change in HCV RNA copies per ml) (Fig. 1a). However, the distribution of V into infectious and non-infectious compartments differs, such that for $\delta = 0.14 \text{ d}^{-1}$ nearly half the virions are non-infectious by day 20 when $\rho_{\max} = 0.5$ (Fig. 1). When ε is low, ribavirin has a noticeable effect on viral load decay. After two months with $\varepsilon = 0.5$, V decays 2 logs and 4 logs for $\rho_{\max} = 0$ and 1, respectively, for $\delta = 0.14 \text{ d}^{-1}$, and the difference increases with δ (Fig. 1b). Further, this difference arises almost entirely from the second-phase decay of V .

In the second phase, viral production and clearance are in pseudo-equilibrium²²; that is, $(1 - \rho)(1 - \varepsilon)pI \approx cV_1$ and $\rho(1 - \varepsilon)pI \approx cV_{NI}$, which yields (Supplementary Note S1)

$$V = (1 - \varepsilon)V_0 \exp[-\delta(\varepsilon + \rho - \varepsilon\rho)t] \quad (4)$$

where V_0 is the viral load at the onset of therapy. Thus, the second-phase slope under combination therapy is $\delta(\varepsilon + \rho - \varepsilon\rho)$. The difference between this slope and the slope under interferon monotherapy ($\rho = 0$) is $\Delta = \delta(\varepsilon + \rho - \varepsilon\rho) - \delta\varepsilon = \delta\rho(1 - \varepsilon)$. Because $\Delta \approx 0$ when $\varepsilon \approx 1$, ribavirin has a negligible effect on the second-phase slope when interferon effectiveness is high. (However, a fraction, ρ , of the virions are non-infectious because of ribavirin; Supplementary Note S2.) As ε decreases, Δ increases and ribavirin enhances the second-phase slope. Viral load decrease in the first phase is about εV_0 and is independent of ρ . Indeed, in a recent study with high-dose daily interferon, in which for white genotype-1-infected subjects the mean ε was estimated at about 0.98, ribavirin seemed to have no effect on first-phase or second-phase kinetics²³.

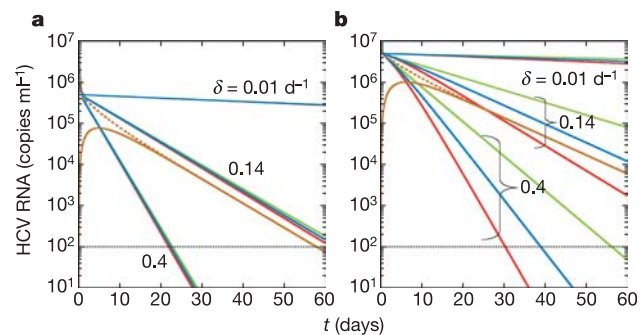


Figure 1 Theoretical viral load decay profiles. Equations (1)–(3) are solved for two values of interferon effectiveness, $\varepsilon = 0.95$ (a) and $\varepsilon = 0.5$ (b), with maximum ribavirin effectiveness of $\rho_{\max} = 1$ (red lines), 0.5 (blue lines) and 0 (green lines), a virion clearance rate of $c = 6.2 \text{ d}^{-1}$, an initial viral load of $V_0 = 10^7 \text{ ml}^{-1}$, and different values of the loss rate of productively infected cells, δ . We assume that the number density of target cells, T , remains constant at the pretreatment value $\delta c/p\beta$ (ref. 22). (Successful therapy should cause T to increase. We have performed calculations where we let T increase twofold over the duration of therapy and found a negligible influence on V .) Choosing a value for the virion production rate p is avoided because multiplication of equation (1) by p when $T = \delta c/p\beta$ enables consideration throughout of the composite variable Ip with initial value cV_0 . The distribution of viral load into infectious (V_1) (dashed orange lines) and non-infectious (V_{NI}) (solid orange lines) compartments is shown for $\delta = 0.14 \text{ d}^{-1}$ and $\rho_{\max} = 0.5$. The detection limit (dotted black line) is 100 copies ml^{-1} .

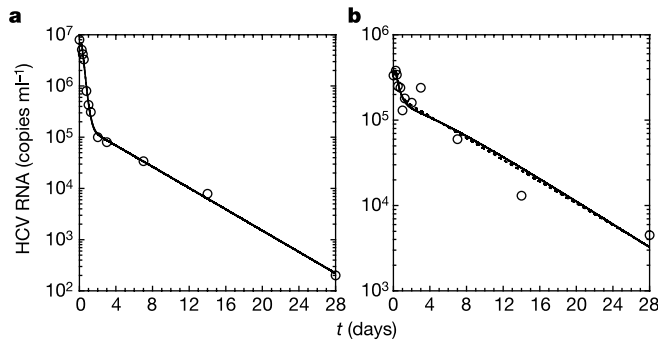


Figure 2 Comparison of theoretical viral load decay profiles with patient data. Best-fit model predictions (representative) of viral load decay with $\rho_{\max} = 1$ (solid lines) and $\rho_{\max} = 0$ (dashed lines) compared with data (symbols) from patients 18 (**a**) and 35 (**b**)²³. Parameters employed (Supplementary Table S1) were $V_0 = 8 \times 10^6 \text{ ml}^{-1}$, $c = 3.8 \text{ d}^{-1}$, $\delta = 0.24 \text{ d}^{-1}$, $\varepsilon = 0.982$ and $\tau = 4.3 \text{ h}$ for both $\rho_{\max} = 1$ and $\rho_{\max} = 0$ (**a**) and $V_0 = 3.6 \times 10^5 \text{ ml}^{-1}$, $c = 2.2 \text{ d}^{-1}$, $\delta = 0.16 \text{ d}^{-1}$, $\varepsilon = 0.6$ and $\tau = 6.2 \text{ h}$ for $\rho_{\max} = 1$, and $V_0 = 3.5 \times 10^5 \text{ ml}^{-1}$, $c = 3.1 \text{ d}^{-1}$, $\delta = 0.31 \text{ d}^{-1}$, $\varepsilon = 0.5$ and $\tau = 7.2 \text{ h}$ for $\rho_{\max} = 0$ (**b**). Note that the two fits overlap in **a**.

At the same time, another study found that ribavirin addition when $\varepsilon \approx 0.5$ had little influence on ε (obtained from the first phase) but increased the average value of δ (obtained from later phase(s)) twofold over interferon monotherapy²⁴. In a more recent study as well, ribavirin seemed to enhance the second-phase slope with standard thrice-a-week interferon in a dose-dependent manner but not with high-dose daily interferon¹². Our model thus reconciles these seemingly conflicting observations that ribavirin addition enhances viral load decline in some cases but not in others.

The above analysis rules out a major antiviral role for the immune modulatory effect of ribavirin, suggested as a key alternative mechanism of ribavirin action against HCV⁸. An immune modulatory effect would enhance the second-phase slope by increasing δ independently of ε . Thus, ribavirin would seem equally potent regardless of interferon effectiveness, contrary to the above observations^{12,23,24}.

We have applied our model to an analysis of viral load data obtained from 17 genotype-1-infected patients under combination therapy, who formed part of a larger study partly aimed at explaining the differences in the responses of white and African American patients to interferon-based therapy²³. Our model provides excellent fits to the data (Fig. 2, Supplementary Table S1 and Supplementary Fig. S1). We find that the best-fit parameter values obtained for patients with high interferon effectiveness are in close agreement with values obtained by models that ignore ribavirin action²³. This agreement is expected because with high interferon effectiveness ($\varepsilon \approx 1$) the addition of ribavirin has little influence on viral load decay. For African American patients in whom ε is much smaller than 1, the average value of δ , the loss rate of productively infected cells, obtained from the present fits is lower than that obtained with models that ignore ribavirin action²³. By

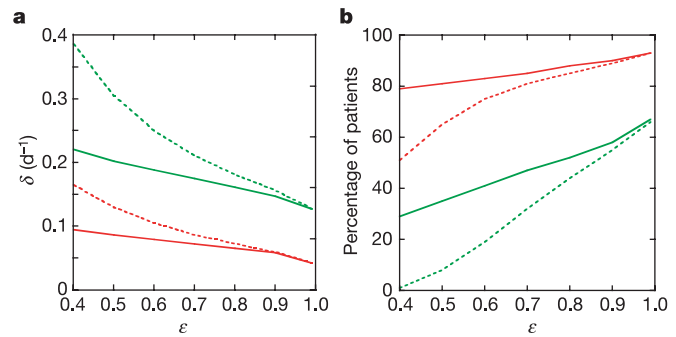


Figure 3 Model predictions of ETR and SVR. **a**, Minimum values of the loss rate of productively infected cells, δ , required for attaining ETR (red) and SVR (green). **b**, Fractions of patients treated for 24 weeks that would exhibit ETR (red) or SVR (green), respectively, with interferon monotherapy (dotted lines) or combination therapy assuming a ribavirin effectiveness of $\rho_{\max} = 0.5$ (solid lines), as functions of interferon effectiveness, ε (see Supplementary Table S2). Parameter values used are the initial viral load, $V_0 = 10^7$ copies ml^{-1} , virion clearance rate, $c = 9.5 \text{ d}^{-1}$, and the duration of the shoulder phase, $\tau = 0.3 \text{ d}$.

assuming a ribavirin effectiveness of $\rho = 0$, the latter models overestimate δ . Further, the average δ from present fits is similar for African American and white patients, indicating that the loss rate of productively infected cells might be race independent. Ribavirin, by lowering the infectivity of virions, diminishes the production rate of infected cells but does not influence their loss rate. That our model is able to distinguish between these subtle effects is remarkable and argues further against an immune modulatory role for ribavirin.

Two measures of the long-term outcome of therapy are commonly employed, namely end-of-treatment response (ETR) and sustained virological response (SVR), defined respectively as the plasma viral load below the assay detection limit at the end of treatment and the plasma viral load remaining below detection 24 weeks after the cessation of therapy. Ribavirin addition substantially increases both ETR and SVR, which we explain quantitatively with our model.

From equation (4), for given V_0 , ε and ρ , a minimum value of the loss rate of infected cells $\delta = \delta_{\text{ETR}}$ is required for V to reach the detection limit of 100 copies ml^{-1} at the end of treatment. A corresponding value, $\delta = \delta_{\text{SVR}}$, is required to reach the clearance limit, which we define as V less than one virion in the typical 15 l extracellular fluid volume (that is, 6×10^{-5} copies ml^{-1}) in humans. We assume that attaining the clearance limit results in SVR. For known ranges of V_0 , ε and ρ , we solve equations (1)–(3) to determine δ_{ETR} and δ_{SVR} (Fig. 3a, and Supplementary Table S2). Patients with $\delta \geq \delta_{\text{ETR}}$ will exhibit ETR and those with $\delta \geq \delta_{\text{SVR}}$ will exhibit SVR. δ varies significantly across patients and typically lies in the range 0.01–0.4 d^{-1} (refs 22–24). Here we assume that δ is normally distributed, with mean $\mu = 0.16 \text{ d}^{-1}$ and standard deviation $\sigma = 0.10 \text{ d}^{-1}$ deduced from fitting patient data (Supplementary Table S1). The probability that $\delta \geq \delta_{\text{ETR}}$ (or δ_{SVR}) then

Table 1 Comparison of model predictions and experimental observations

Duration of therapy		ETR (%)		SVR (%)	
		Interferon	Interferon + ribavirin	Interferon	Interferon + ribavirin
24 weeks	Observation	37–59	64–77	8–21	39–53
	Model prediction	65	80	*	35–41
48 weeks	Observation	31–68	71–94	17–24	53–72
	Observation (Peg-Interferon)	91	93–100	45	77
	Model prediction	88	93	56	76

The ETR and SVR are for interferon monotherapy and combination therapy with ribavirin (see refs 4, 13–15 and Supplementary Table S3).

*We employ observed SVR for 24 weeks of interferon monotherapy to set the interferon efficacy, ε , and thus do not predict it. SVR of 6–18% following 24 weeks of interferon monotherapy has been reported²⁴, which upon discounting for patients who did not complete the treatment course or were under reduced dose yields SVR of 8–21% (Supplementary Table S2). In our calculations, this range corresponds to $\varepsilon \sim 0.5$ – 0.6 (Fig. 3), which is within the range for ε determined from viral dynamics studies²⁴. Addition of ribavirin when $\varepsilon \sim 0.5$ – 0.6 yields an SVR prediction of 35–41% (Fig. 3, Supplementary Table S2) in agreement with the experimentally observed range of 39–53% (Supplementary Table S3). The remaining model predictions are all made with $\varepsilon \sim 0.5$.

yields the fraction of patients that would exhibit ETR (or SVR) (Fig. 3b and Supplementary Table S2). The best fits to the ETR and SVR data are obtained by assuming that 1,200 mg of ribavirin daily when given with interferon corresponds to $\rho_{\max} = 0.5$ (Supplementary Note S3). The difference in the ETR (SVR) values for $\rho_{\max} = 0$ and $\rho_{\max} = 0.5$ quantifies the enhancement in ETR (SVR) due to ribavirin addition.

Model predictions of ETR and SVR are in good agreement with experiments (Table 1). With $\varepsilon \approx 0.5$, obtained by fitting model predictions to SVR after 24 weeks of standard thrice-a-week interferon monotherapy, our model captures ETR and SVR after 24 weeks of standard interferon therapy and 48 weeks of pegylated interferon therapy with and without ribavirin. After 48 weeks of standard interferon therapy, however, model predictions overestimate ETR and SVR, indicating that $\varepsilon < 0.5$ for this case. Why ε seems to decrease from 24 to 48 weeks of standard interferon remains unknown. Perhaps the consistently higher incidence of side effects for 48 weeks of therapy³ compromises patient compliance. Greater compliance might be achieved with pegylated interferon, given the less frequent, weekly dosing, and might contribute to higher response rates (Supplementary Note S4). From all studies, the maximum enhancement in ETR or SVR by ribavirin addition is about 25–30% (Supplementary Table S3), as predicted by our model (Supplementary Table S2).

Our model thus provides a convincing picture of how ribavirin enhances HCV RNA decline and improves the long-term outcome of interferon-based therapy. Significant clinical implications follow. Because ribavirin influences the second and not the first phase of viral load decline, ε determined from the first phase underpredicts the long-term outcome of combination therapy²⁶. Because of enhanced second-phase decline, patients with slightly slower decays who might not achieve viral negativity under interferon monotherapy might become responders with combination therapy²⁷. In patients who achieve poor interferon effectiveness, for example some African American patients²³ or some patients infected with HCV genotype 1 (ref. 28), ribavirin addition should enhance second-phase decay²⁹ and improve ETR and SVR. For pegylated interferon α -2b administered weekly, during which ε might decrease significantly between doses³⁰, combination therapy might be particularly desirable. When ε is high, viral production is suppressed by interferon. As ε drops between doses, ribavirin can decrease the infectivity of new virions and prevent viral load resurgence¹².

By explicitly incorporating the anti-HCV activity of ribavirin, which extant models ignore^{22–24}, our model establishes a much-needed framework for the rational optimization of therapy. However, exploiting this framework requires an understanding of the possible synergy between ribavirin and interferon. This synergy is suggested by the observations that at the same ribavirin dosage, low ribavirin efficacy, $\rho_{\max} \ll 1$, accounts for the observed insignificant viral load decline under ribavirin monotherapy (Supplementary Note S5), but a much higher efficacy, $\rho_{\max} \approx 0.5$, explains long-term responses with combination therapy. We speculate on how this synergy might occur based on the modes of ribavirin and interferon action against HCV. Ribavirin is incorporated into the RNA of replicating genomes and gives rise to mutations, whose inherent frequency might be low. In the presence of interferon, viral production is diminished. The relative ribavirin concentration per replicating genome at membranous replication sites in an infected cell might then be elevated, which could increase the mutation frequency. At the same time, ribavirin lowers the intracellular guanosine triphosphate pools⁸, which might further increase the likelihood of ribavirin incorporation. When the mutation frequency is sufficiently high, replicating virions become non-infectious. Interferon could therefore provide the necessary impetus for the mutagenic activity of ribavirin to exert an observable antiviral effect against HCV. Although recent measurements *in vivo* indicate that mutation frequencies might not be significantly enhanced under

ribavirin monotherapy¹², mutation frequencies under combination therapy, which provide tests of this hypothesis, have not been measured. Nevertheless, our model applies so long as ribavirin decreases HCV infectivity, even if this occurs by mechanisms other than mutagenesis. □

Received 10 August; accepted 25 October 2004; doi:10.1038/nature03153.

1. World Health Organization, Hepatitis C—global prevalence (update). *World Health Org. Weekly Epidemiol. Rec.* **75**, 18–19 (2000).
2. National Institutes of Health. Consensus Statement on Management of Hepatitis C. *NIH Consens. State Sci. Statements* **19**, 1–46 (2002).
3. Manns, M. P. *et al.* Peginterferon α -2b plus ribavirin compared with interferon α -2b plus ribavirin for initial treatment of chronic hepatitis C: a randomised trial. *Lancet* **358**, 958–965 (2001).
4. Fried, M. W. *et al.* Peginterferon α -2a plus ribavirin for chronic hepatitis C virus infection. *N. Engl. J. Med.* **347**, 975–982 (2002).
5. Tan, S.-L., Pause, A., Shi, Y. & Sonenberg, N. Hepatitis C therapeutics: Current status and emerging strategies. *Nature Rev. Drug Discov.* **1**, 867–881 (2002).
6. Shiffman, M. L. *et al.* A randomized, controlled trial to determine whether continued ribavirin monotherapy in hepatitis C virus-infected patients who responded to interferon-ribavirin combination therapy will enhance sustained virologic response. *J. Infect. Dis.* **184**, 405–409 (2001).
7. Hoofnagle, J. H. *et al.* Maintenance therapy with ribavirin in patients with chronic hepatitis C who fail to respond to combination therapy with interferon alpha and ribavirin. *Hepatology* **38**, 66–74 (2003).
8. Lau, J. Y. N., Tam, R. C., Liang, T. J. & Hong, Z. Mechanism of action of ribavirin in the combination treatment of chronic HCV infection. *Hepatology* **35**, 1002–1009 (2002).
9. Dusheiko, G. *et al.* Ribavirin treatment for patients with chronic hepatitis C: results of a placebo-controlled study. *J. Hepatol.* **25**, 591–598 (1996).
10. Bodenheimer, H. C. *et al.* Tolerance and efficacy of oral ribavirin treatment of chronic hepatitis C: A multicenter trial. *Hepatology* **26**, 473–477 (1997).
11. Zoulim, F. *et al.* Ribavirin monotherapy in patients with chronic hepatitis C: a retrospective study of 95 patients. *J. Viral Hepat.* **5**, 193–198 (1998).
12. Pawlotsky, J.-M. *et al.* Antiviral action of ribavirin in chronic hepatitis C. *Gastroenterology* **126**, 703–714 (2004).
13. McHutchison, J. G. *et al.* Interferon α -2b alone or in combination with ribavirin as initial treatment for chronic hepatitis C. *N. Engl. J. Med.* **339**, 1485–1492 (1998).
14. Reichard, O. *et al.* Randomized, double-blind, placebo-controlled trial of interferon α -2b with and without ribavirin for chronic hepatitis C. *Lancet* **351**, 83–87 (1998).
15. Poynard, T. *et al.* Randomized trial of interferon α 2b plus ribavirin for 48 weeks or for 24 weeks versus interferon α 2b plus placebo for 48 weeks for treatment of chronic infection with hepatitis C virus. *Lancet* **352**, 1426–1432 (1998).
16. Lanford, R. E. *et al.* Antiviral effect and virus-host interactions in response to alpha interferon, gamma interferon, poly(I)-poly(C), tumor necrosis factor alpha, and ribavirin in hepatitis C virus subgenomic replicons. *J. Virol.* **77**, 1092–1104 (2003).
17. Contreras, A. M., Viral, R. N. A. *et al.* mutations are region specific and increased by ribavirin in a full-length hepatitis C virus replication system. *J. Virol.* **76**, 8505–8517 (2002).
18. Young, K.-C. *et al.* Identification of a ribavirin-resistant NS5B mutation of hepatitis C virus during ribavirin monotherapy. *Hepatology* **38**, 869–878 (2003).
19. Lanford, R. E. *et al.* Ribavirin induces error-prone replication of GB virus B in primary tamarin hepatocytes. *J. Virol.* **75**, 8074–8081 (2001).
20. Hong, Z. The role of ribavirin-induced mutagenesis in HCV therapy: A concept or a fact? *Hepatology* **38**, 807–810 (2003).
21. Crotty, S., Cameron, C. E. & Andino, R. RNA virus error catastrophe: Direct molecular test by using ribavirin. *Proc. Natl Acad. Sci. USA* **98**, 6895–6900 (2001).
22. Neumann, A. U. *et al.* Hepatitis C viral dynamics *in vivo* and the antiviral efficacy of interferon- α therapy. *Science* **282**, 103–107 (1998).
23. Layden-Almer, J. E., Ribeiro, R. M., Wiley, T., Perelson, A. S. & Layden, T. J. Viral dynamics and response differences in HCV-infected African American and white patients treated with IFN and ribavirin. *Hepatology* **37**, 1343–1350 (2003).
24. Herrmann, E., Lee, J.-H., Marison, G., Modi, M. & Zeuzem, S. Effect of ribavirin on hepatitis C virus kinetics in patients treated with pegylated interferon. *Hepatology* **37**, 1351–1358 (2003).
25. Glue, P. The clinical pharmacology of ribavirin. *Semin. Liv. Dis.* **19**, 17–24 (1999).
26. Steindl-Munda, P. *et al.* in *Abstracts 11th Int. Symp. Viral Hepatitis and Liver Dis.* 75–76 (2003).
27. Tsubota, A. *et al.* Viral dynamics and pharmacokinetics in combined interferon α -2b and ribavirin therapy for patients infected with hepatitis C virus of genotype 1b and high pretreatment viral load. *Intervirology* **45**, 33–42 (2002).
28. Neumann, A. U. *et al.* Differences in viral dynamics between genotypes 1 and 2 of hepatitis C virus. *J. Infect. Dis.* **182**, 28–35 (2000).
29. McHutchison, J. G. *et al.* The impact of interferon plus ribavirin on response to therapy in black patients with chronic hepatitis C. *Gastroenterology* **119**, 1317–1323 (2000).
30. Powers, K. A. *et al.* Modeling viral and drug kinetics: hepatitis C virus treatment with pegylated interferon α -2b. *Semin. Liv. Dis.* **23**, 13–18 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank R. M. Ribeiro, J.-M. Pawlotsky, S. Zeuzem, X. Tong and B. Malcolm for helpful comments. Portions of this work were performed under the auspices of the US Department of Energy and supported by grants from the NIH, the University of Illinois and the Chicago VA.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.S.P. (asp@lanl.gov).

Substrate recognition strategy for botulinum neurotoxin serotype A

Mark A. Breidenbach¹ & Axel T. Brunger²

¹Department of Molecular and Cellular Physiology and ²Howard Hughes Medical Institute and Departments of Molecular and Cellular Physiology, Neurology and Neurological Sciences and Stanford Synchrotron Radiation Laboratory, Stanford University, Stanford, California 94305, USA

Clostridal neurotoxins (CNTs) are the causative agents of the neuroparalytic diseases botulism and tetanus^{1,2}. CNTs impair neuronal exocytosis through specific proteolysis of essential proteins called SNAREs³. SNARE assembly into a low-energy ternary complex is believed to catalyse membrane fusion, precipitating neurotransmitter release; this process is attenuated in response to SNARE proteolysis^{4–7}. Site-specific SNARE hydrolysis is catalysed by the CNT light chains, a unique group of zinc-dependent endopeptidases³. The means by which a CNT properly identifies and cleaves its target SNARE has been a subject of much speculation; it is thought to use one or more regions of enzyme–substrate interaction remote from the active site (exosites)^{8–10}. Here we report the first structure of a CNT endopeptidase in complex with its target SNARE at a resolution of 2.1 Å: botulinum neurotoxin serotype A (BoNT/A) protease bound to human SNAP-25. The structure, together with enzyme kinetic data, reveals an array of exosites that determine substrate specificity. Substrate orientation is similar to that of the general zinc-dependent metalloprotease thermolysin¹¹. We observe significant structural changes near the toxin's catalytic pocket upon substrate binding, probably serving to render the protease competent for catalysis. The novel structures of the substrate-recognition exosites could be used for designing inhibitors specific to BoNT/A.

The structures of three CNT light chains suggest that substrate recognition cannot occur at the active sites of these proteases because the catalytic pocket composition and geometries of BoNTs A, B and E are essentially identical^{12–14} (Fig. 1). Furthermore (and atypically for endopeptidases), light-chain activity can be strongly influenced by remote substitutions and deletions^{10,15,16}. For example, conserved motifs containing acidic residues in the substrates were shown to be required for normal levels of light-chain activity and led to the proposal that the light chains may use exosites for efficient substrate recognition and cleavage^{8–10,17}. However, structural data concerning the locations and functions of these exosites have remained elusive until now.

We created an inactive variant of BoNT/A light chain, because attempts to co-crystallize the wild-type BoNT/A light chain failed. Although a single-point mutation (E224Q) substantially impaired substrate turnover as previously observed¹⁸, we detected cleavage products at the high concentrations of enzyme required for crystallization. A second mutation known to impair catalysis, Y366F, was added to eliminate substrate turnover at the conditions required for crystallization¹⁹. As a control, the double-mutant apo structure was solved by molecular replacement to 2.2 Å resolution (see Supplementary Methods) and found to be essentially identical to the wild-type structure¹², including side-chain orientations at positions 224 and 366 (Fig. 1). The carboxy-terminal SNARE (soluble N-ethylmaleimide-sensitive factor attachment protein receptor) domain of synaptosome-associated protein-25 kDa (SNAP-25), residues 141–204 (referred to as sn2), was co-crystallized with the double mutant BoNT/A light chain, and the structure of the complex was solved by molecular replacement to 2.1 Å resolution (see Supplementary Methods). Electron density for nearly all of sn2 (residues 146–204) was observed and independently confirmed by

experimental anomalous sulphur phasing. The interface between the endopeptidase and substrate is extensive, wrapping around most of the light chain's circumference (Fig. 2). The interface buries approximately 4,840 Å² of surface area, about three times the area observed in thrombin–inhibitor complexes²⁰. In contrast to the contiguous helical conformation of sn2 within the ternary SNARE core complex²¹, sn2 adopts three distinct types of secondary structure in complex with BoNT/A: residues 147–167 form a distorted α-helix, residues 168–200 are extended and residues 201–204 are involved in a distorted β-sheet conformation. The observed interface between BoNT/A and sn2 is consistent with the previous report of a truncated substrate that is cut with the same efficiency as full-length SNAP-25 (ref. 15). The validity of a previously reported BoNT/B–SNARE complex¹³ has been questioned²², primarily owing to the absence of observed electron density for the substrate (see Supplementary Fig. 2). The reported BoNT/B–SNARE complex also features highly improbable ligand stereochemistry and provides little insight regarding determinants of substrate specificity, making our sn2–BoNT/A complex the first credible structure of a substrate-bound CNT light chain.

The amino-terminal helical region of sn2 interacts with the endopeptidase along a hydrophobic patch formed at the interface of four light-chain α-helices, referred to as the α-exosite (Figs 2, 3a): residues 102–113 (α-helix 1), 310–321 (α-helix 2), 335–348 (α-helix 3), and 351–358 (α-helix 4). The hydrophilic face of the amphipathic sn2 helix is oriented towards the solvent, while the hydrophobic side chains are buried against the α-exosite. Mutations to hydrophobic sn2 residues at this interface, including I156E and M167E, substantially decrease k_{cat}/K_m ratios (catalytic efficiency) compared to wild-type sn2 by increasing K_m (the Michaelis constant) (Fig. 4). N-terminally truncated substrates incapable of binding the α-exosite exhibit a further reduction in k_{cat}/K_m ; the reported K_m for SNAP-25 residues 187–203 is ~55 times greater than that for SNAP-25 residues 1–206, whereas reported k_{cat} (turnover rate) values for these substrates are identical within

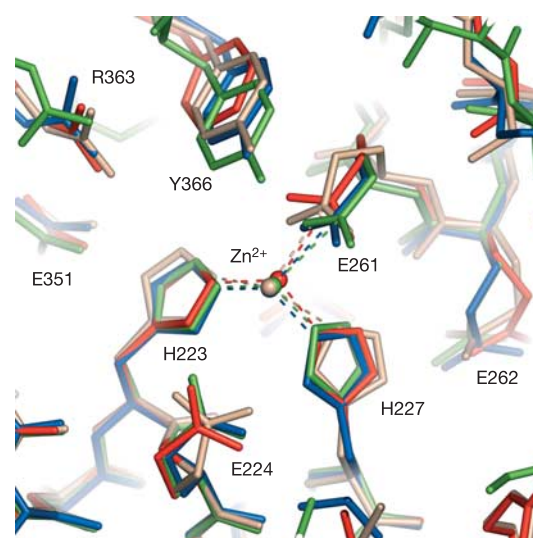


Figure 1 The nearly identical active sites of different BoNT light chains. The superimposed catalytic pockets of BoNT/A (tan), the E224Q/Y366F mutant of apo BoNT/A (red), BoNT/B (blue), and BoNT/E (green), with target sites P₄P₃P₂P₁/P₁'P₂'P₃'P₄' (that is, the residue positions upstream and downstream of the protease cleavage site) Glu-Ala-Asn-Gln/Arg-Ala-Thr-Lys (BoNT/A), Gly-Ala-Ser-Gln/Phe-Glu-Thr-Ser (BoNT/B) and Gln-Ile-Asp-Arg/Ile-Met-Glu-Lys (BoNT/E) shown with BoNT/A residue numbering. Coordinates for wild-type BoNT/A, BoNT/B, and BoNT/E are from the Protein Data Bank (accession codes 3BTA, 1F82, 1T3A)^{12–14}. Notably, the structure of the E224Q/Y366F mutant of BoNT/A is similar to that of wild-type BoNT/A. The fourth ligand of Zn²⁺ is a water molecule coordinated by Glu 224; for clarity it is not shown.

the margins of error^{18,23,24}. Interestingly, conserved acidic residues of sn2 including Asp 147, Glu 148, and Glu 151 do not directly contact BoNT/A. Similar groups of charged residues have been implicated in mediating light chain–SNARE interactions in BoNTs B, G and tetanus neurotoxin, suggesting that SNARE recognition sites differ among the light chains⁹. Notably, the α -exosite of the BoNT/A light chain is structurally distinct from the hydrophobic interface formed in the ternary SNARE core complex²¹ that binds the same stretch of sn2.

In the BoNT/A light chain–sn2 complex, the C-terminal portion of sn2 forms one strand of a distorted, three-stranded antiparallel β -sheet in a region that we refer to as the β -exosite (Figs 2, 3b, 5). This is the first example of β -sheet conformation observed in any SNARE protein, and further illustrates the conformational variability of SNAREs²⁵. The other two strands of the β -sheet are contributed by the '250 loop' of the light chain, a region encompassing residues 242–259 (Figs 2, 3b). Specifically, residues 252–257 of the central strand pair with sn2 residues 201–204. The β -exosite interface is primarily mediated by backbone interactions, with the exception of Met 202 of sn2. Met 202 packs into a hydrophobic cavity formed by the interface of the light-chain 250 loop and the '370 loop', a region that encompasses light-chain residues 359–370. The 370 loop protrudes into the active site (Figs 2, 5) and separates the 250 loop from the catalytic pocket. Kinetic data linking the β -exosite (including the 370 loop and the C terminus of the substrate) to k_{cat} are available: the BoNT/A Y366F and R363A mutations are known to reduce k_{cat} $\sim$ 40-fold and $\sim$ 80-fold respectively, but K_{m} values remain nearly identical to the wild type¹⁹. Our kinetic characterization of the sn2 mutant M202Y revealed that the increased K_{m} of this mutant substrate was offset by an increase in k_{cat} , providing further evidence that structural changes in the β -exosite can directly influence k_{cat} (Fig. 4). Notably, substrates with C-terminal truncations at residue positions lower than Met 202 cannot be cut by BoNT/A at a measurable rate²⁴.

Connecting the α - and β -exosites, the extended portion of sn2 is

channelled around the BoNT/A surface, leading directly through the catalytic pocket (Fig. 2). A small portion of the sn2 chain (residues 183–190) detaches from the surface of the light chain, rejoining the endopeptidase at sn2 residue Arg 191 and leading directly into the active site. Substrate electron density in the catalytic pocket (from sn2 residues 197–199) is fragmented, probably owing to the BoNT/A Y366F mutation, which might prevent a critical enzyme–substrate contact. Electron density is sufficient, however, to show a sharp, stereochemically strained turn leading out of the active site and connecting directly to the β -exosite (Fig. 5).

Several residues in the extended portion of sn2 are involved in side chain–side chain contacts with the endopeptidase at a series of anchor points that could serve as additional determinants of substrate specificity (Fig. 2, Supplementary Fig. 1). Substrate residues 170–172 and 192–193 comprise two distinct, contiguous anchor points, and individual side chains from Arg 176, Ile 178, Ile 181 and Glu 183 are also recognized by the endopeptidase. These anchor points, accounting for nearly half of the side chain–side chain contacts, assist in directing the path of the substrate into the endopeptidase's active site and can be classified as additional exosites (see Supplementary Fig. 1).

The availability of multiple apo crystal structures^{12,26} along with our structure of the complex allowed us to discriminate between conformational changes resulting from crystal contacts, substrate binding and point mutations. Along the interface spanning from the α -exosite to the catalytic pocket, we observe only minor side-chain rearrangements in the light chain. The largest structural changes that are related to sn2 binding occur in the 250 loop and 370 loop at the β -exosite (Fig. 5). In apo BoNT/A, the 250 loop adopts an open conformation which contracts and folds over the 370 loop upon sn2 binding at the β -exosite (Fig. 5). Bulky side chains from both sn2 and the 250 loop including Met 202 and Tyr 250, respectively, pack against a stretch of the 370 loop and also influence its conformation.

Additional changes stem from the interaction of the Y366F

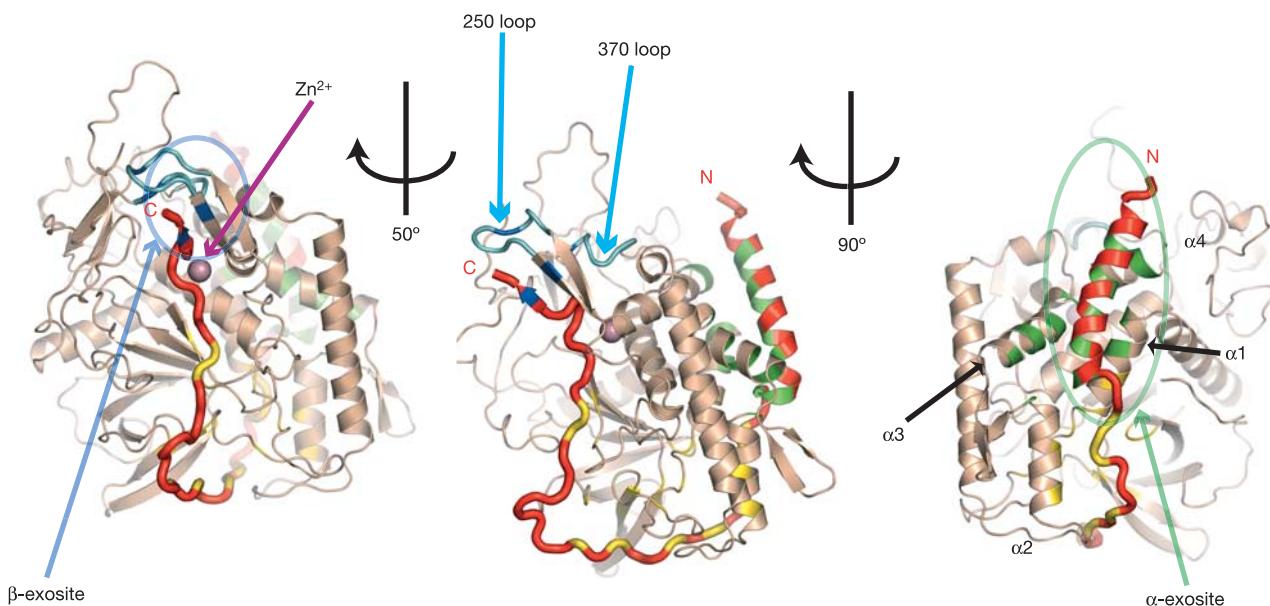


Figure 2 The interface between sn2 and BoNT/A. Three different views of the sn2–BoNT/A complex are shown; these are related by the specified rotations around a vertical axis in the plane of the figure that goes through the centre of the complex. The α -exosite (indicated by a green arrow) is formed by BoNT/A light-chain helices (tan) α 1– α 4 that bind to the helical N terminus of sn2 (red). Green areas indicate the approximate locations of contacting side chains involved in the α -exosite. On the opposite face of sn2–BoNT/A, the β -exosite is indicated by a blue arrow. The C terminus of sn2 forms an anti-parallel β

sheet along with a portion of the 250 loop (light-blue), which is separated from the active site (indicated by Zn^{2+} , purple sphere) by the 370 loop (light-blue). Dark-blue areas indicate the approximate location of contacting side chains involved in the β -exosite. Yellow areas indicate the approximate location of other residues (anchor points) involved in side-chain contacts between the sn2 substrate and the BoNT/A light chain. A detailed list of all contacts between sn2 and BoNT/A light chain can be found in Supplementary Fig. 1.

mutation with sn2. Incapable of forming a hydrogen bond with the substrate, the mutant Phe 366 residue is reoriented to point away from sn2 (Figs 1, 5). The new Phe 366 position causes a cascade of additional side-chain reorientations, leading to salt-bridge formation between Arg 363 and Asp 370 (Fig. 5). This salt bridge buttresses the entire 370 loop. We speculate that in wild-type BoNT/A, Tyr 366 would not be redirected away from Arg 198 and the salt bridge between Arg 363 and Asp 370 would not form.

BoNT/A shares the same His-Glu-X-X-His zinc-binding motif as thermolysin^{12,19}. Our structure reveals that the directionality of the substrate is the same as that of peptide inhibitors bound to thermolysin¹¹ and that the positions of the primary catalytic

residues match those found in thermolysin (see Supplementary Fig. 3). The scissile peptide bond of sn2, between Gln 197 and Arg 198, is positioned near the catalytic water molecule that is coordinated by Zn^{2+} (Fig. 5). Consistent with this model, the BoNT/A E224Q mutant no longer efficiently deprotonates the intervening water molecule. However, there are no obvious residues similar to thermolysin Tyr 157 and His 231, which have been proposed to stabilize the transition state²⁷. In fact, the space occupied by these two residues in thermolysin is occupied by sn2 residues in the case of BoNT/A. Thus, it is possible that there are no specific residues to stabilize the evolving oxyanion, or that BoNT/A activity might involve substrate-assisted catalysis²⁸, where polar groups at the sharp kink following the scissile bond (such as the side-chain hydroxyl or backbone amide of Thr 200) could stabilize the transition state.

Interestingly, a homodimeric structure of BoNT/A shows an interaction between a C-terminal fragment of the autolysed 250

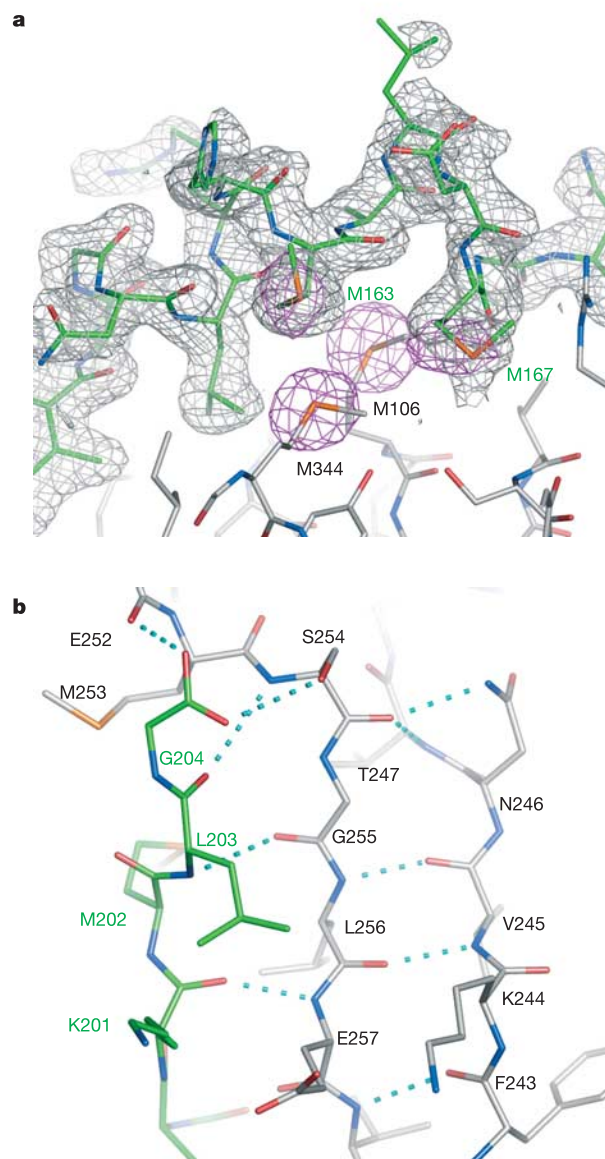


Figure 3 Detailed views of α - and β -exosites. **a**, A cross-validated, σ_A -weighted, phase-combined $(2m|F_o| - D|F_c|)\phi_{comb}$, where ϕ_{comb} and ϕ_{calc} are the combined and model phases, respectively, m is the combined figure of merit and D is an estimate of the model's completeness, both obtained from the σ_A method) spherical OMIT map with simulated annealing is shown, centred on Met 163 and contoured at 1σ (grey mesh). Anomalous Fourier difference density (violet mesh) contoured at 4σ highlights sulphur atoms in sn2 and nearby BoNT/A methionine (M) residues. **b**, The β -sheet network of hydrogen bonds (cyan dashes) between the 250 loop (grey backbone) and sn2 (green backbone) is shown.

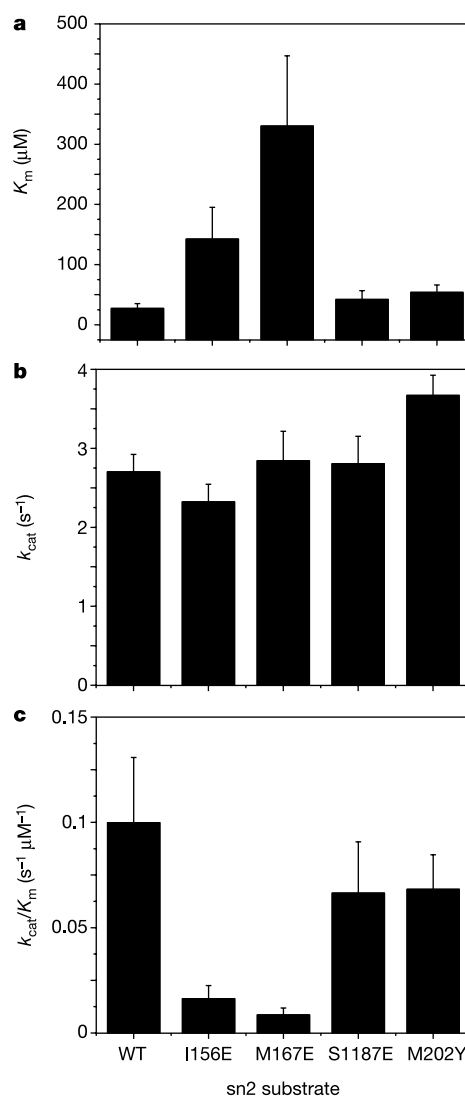


Figure 4 Kinetic characterization of sn2 mutants. Kinetic analyses (see Supplementary Methods) were conducted for mutant substrates that were designed to disrupt hydrophobic packing observed at the α - and β -exosites. S1187E serves as a negative control because it makes no contacts with the light chain in our structure. **a**, K_m values for wild-type (WT) and mutant substrates. **b**, Values for k_{cat} for wild-type and mutant substrates. **c**, The ratio k_{cat}/K_m for each substrate. Error bars represent estimates of s.e.m. Kinetic parameters obtained using the wild-type sn2 substrate are consistent with previously reported results¹⁸.

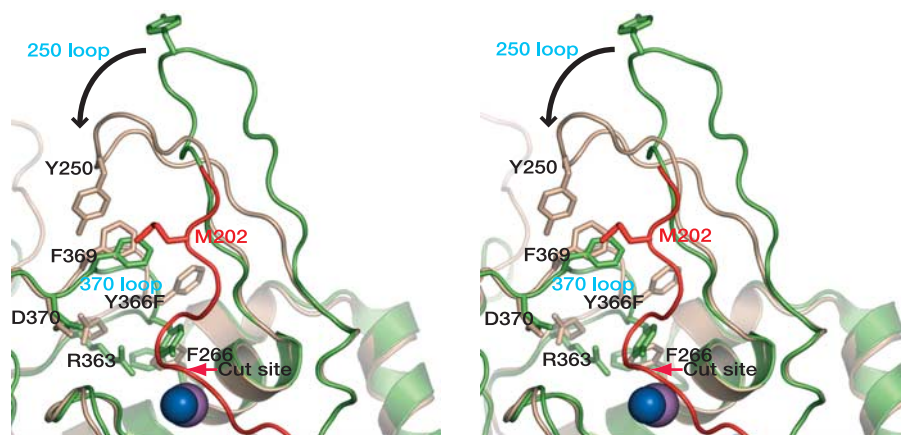


Figure 5 Wall-eyed stereo view of β -exosite conformational changes. The 250 loop switches from an open (green) to a closed (tan) conformation upon sn2 (red) binding. With substrate bound, the 250 loop packs against the 370 loop, which borders the catalytic pocket. The scissile bond (indicated by red arrow) is suspended above the nucleophilic water (blue sphere), shown coordinated by Zn^{2+} (purple sphere) in the active site.

Distortion of the 370 loop is restricted by a salt bridge between Arg 363 and Asp 370. A void caused by the reorientation of the mutant Phe 366 is filled by Phe 266, displacing the Arg 363 side chain. In wild-type BoNT/A, we speculate that deformation of a more dynamic 370 loop could place the acyl carbon of Gln 197 within range of the attacking nucleophile.

loop of one monomer with the active site of the other BoNT/A monomer²⁶. However, the C terminus has the opposite (non-canonical) orientation compared to our structure and thermolysin. This structure probably does not represent a physiologically relevant product-bound configuration of the native substrate. We note that BoNT/A is able to cleave non-specific substrates (such as the 250 loop from another BoNT/A monomer) in the non-canonical orientation, but this occurrence seems to be restricted to the conditions of low pH (4.6) and high protein concentration used in the crystallization of this complex. As mentioned above, a purported product-bound BoNT/B structure¹³ has been shown to be an artefact of model bias²² (see Supplementary Fig. 2).

Based on our structure and available kinetic data for several SNAP-25 substrates, a general model of the strategy used by BoNT/A to recognize and cleave SNAP-25 is presented in Fig. 6a–c. With a full-length substrate, sn2 helix induction can occur at the α -exosite, orienting the peptide to assist in the formation of anchor points along the extended substrate chain (Fig. 6b). BoNT/A can cleave small peptides (sn2 residues 192–206) capable of binding only one anchor point in addition to the β -exosite, but only with a substantial reduction in $k_{\text{cat}}/K_{\text{m}}$ (refs 18, 23, 24). The absence of a conformational change in BoNT/A upon binding to the α -exosite and anchor points is consistent with unchanged k_{cat} values for full-length and N-terminally truncated peptides, as well as our I156E

and M167E point mutants. Thus, most of the enzyme–substrate interface serves to provide a substrate-specific boost to $k_{\text{cat}}/K_{\text{m}}$ by reducing K_{m} . The entropic penalty of orienting an otherwise unstructured polypeptide chain^{25,29} into the observed complex may be largely offset in a substrate-specific manner by the unusually large number of substrate–enzyme contacts at the exosites. Although the β -exosite itself does not confer much substrate specificity, its presence is vital for activity²⁴ and it probably serves to activate the endopeptidase via its contacts with the 370 loop. Our structure also explains the inability of BoNT/A to cleave SNAP-25 when it is present as a complex with other SNAREs (Fig. 6d), because the complexed sn2 is restricted to an entirely helical conformation.

The locations and compositions of the exosites are likely to vary between different BoNT light chains. The α -exosite used by BoNT/A is not conserved on the surface of BoNT/E, which also cuts the same sn2 substrate albeit at a different position. The placement of the exosites may be the primary determinants of scissile bond location and might account for the 17-residue shift in cleavage sites between BoNT/A and BoNT/E. Our structural description of the critical exosites on the surface of BoNT/A that recognize sn2 may lead to development of highly specific inhibitors for this neurotoxin. In particular, the α -exosite could be a candidate for structure-based design using a strategy similar to that used in the successful

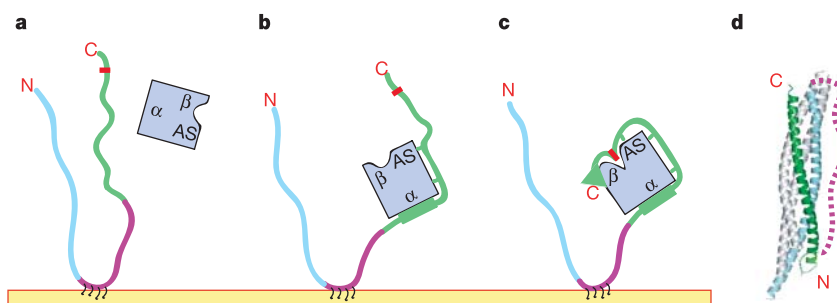


Figure 6 Exosite-based model of BoNT/A substrate recognition. **a**, SNAP-25 is shown attached to a presynaptic membrane via palmitoylation sites (shown in black) on its linker domain (purple). The N-terminal (sn1, cyan) and C-terminal (sn2, green) domains are unstructured or flexible in uncomplexed SNAP-25 (ref. 29). **b**, Binding of BoNT/A (blue) is probably initiated by helix formation at the α -exosite, and anchor points along the extended portion of SNAP-25 (green notches) are additional determinants of substrate

specificity. **c**, These sites reduce K_{m} and enhance binding at the β -exosite, inducing conformational changes at the active site (AS) via the 370 loop, which render the endopeptidase competent to cleave its substrate. **d**, The sn2 domain of SNAP-25 (same colour scheme as in **a–c**) adopts a continuous helical conformation in the SNARE complex.

development of HIV-1 entry inhibitors based on the core structure of gp41 (ref. 30). □

Methods

Details and references are provided in the Supplementary Information which accompanies this paper. Briefly, recombinant apo BoNT/A E224Q/Y366F (residues 2–420) was crystallized by hanging-drop vapour diffusion against 15–18% PEG 3350 and 200 mM di-sodium hydrogen phosphate dihydrate at 20 °C. These crystals indexed in space group *P*2 with near-orthorhombic unit cell dimensions *a* = 57.90 Å, *b* = 40.49 Å, *c* = 195.89 Å and β = 90.25°. Co-crystallization of the complex was performed using a mixture of the inactive BoNT/A light chain with a molar excess of SNAP-25 (residues 141–204), equilibrated against 10% (w/v) PEG 8000, 200 mM magnesium acetate and 100 mM sodium cacodylate (pH 6.5) at 4 °C. These crystals indexed in space group *P*4₃2₁2 with unit cell dimensions *a* = *b* = 86.0 Å, *c* = 165.4 Å. All diffraction data were collected at the Advanced Light Source (beamlines 8.2.1 and 8.2.2) and at the Stanford Synchrotron Radiation Lab (beamline 9-2). The diffraction data were indexed and scaled with MOSFLM and SCALA in the CCP4 computational suite. Initial phases for both structures were obtained by molecular replacement (MR) using light chain coordinates from wild-type BoNT/A (Protein Data Bank accession code 3BTA). For the complex, experimental sulphur single-wavelength anomalous dispersion (SAD) phases were computed and used both as an independent confirmation of the sn2 trace and to supplement the MR phases. All phasing and refinement calculations were performed using the Crystallography and NMR System version 1.1 (CNS). The E224Q/Y366F apo BoNT/A structure was solved to 2.2 Å resolution with final *R*_{working} = 22.0% and *R*_{free} = 27.3%, where *R*_{working} = $\sum |F_o - F_c| / \sum |F_o|$ (where *F*_o and *F*_c are the observed and model structure factor amplitudes, respectively) and *R*_{free} is equivalent to *R*_{working} but is calculated for a randomly chosen 5% of reflections excluded from model refinement. The sn2–BoNT/A complex structure was determined to 2.1 Å with final *R*_{working} = 21.9% and *R*_{free} = 24.9%. Structural illustrations were prepared using Pymol. Kinetic parameters were determined by quantifying C-terminal cleavage products of various sn2 substrates following partial digestion with wild-type BoNT/A light chain. Cleavage products were fractionated by reverse-phase high-performance liquid chromatography (RP-HPLC) on a 218TP54 column (Vydac), and plots of reaction velocity as a function of substrate concentration were fitted to the Michaelis-Menten equation to yield *K*_m, *k*_{cat} and error estimates.

Received 21 September; accepted 19 October 2004; doi:10.1038/nature03123.
Published online 12 December 2004.

- Humeau, Y., Doussau, F., Grant, N. J. & Poulain, B. How botulinum and tetanus neurotoxins block neurotransmitter release. *Biochimie* **82**, 427–446 (2000).
- Dolly, J. O., Black, J., Williams, R. S. & Melling, J. Acceptors for botulinum neurotoxin reside on motor nerve terminals and mediate its internalization. *Nature* **307**, 457–460 (1984).
- Schiavo, G. *et al.* Tetanus and botulinum-B neurotoxins block neurotransmitter release by proteolytic cleavage of synaptobrevin. *Nature* **359**, 832–835 (1992).
- Chen, Y. A., Scales, S. J., Patel, S. M., Doung, Y. C. & Scheller, R. H. SNARE complex formation is triggered by Ca²⁺ and drives membrane fusion. *Cell* **97**, 165–174 (1999).
- Söllner, T. *et al.* SNAP receptors implicated in vesicle targeting and fusion. *Nature* **362**, 318–324 (1993).
- Pellegrini, L. L., O'Connor, V., Lottspeich, F. & Betz, H. Clostridial neurotoxins compromise the stability of a low energy SNARE complex mediating NSF activation of synaptic vesicle fusion. *EMBO J.* **14**, 4705–4713 (1995).
- Xu, T., Binz, T., Niemann, H. & Neher, E. Multiple kinetic components of exocytosis distinguished by neurotoxin sensitivity. *Nature Neurosci.* **1**, 192–200 (1998).
- Rossetto, O. *et al.* SNARE motif and neurotoxins. *Nature* **372**, 415–416 (1994).
- Pellizzari, R. *et al.* Structural determinants of the specificity for synaptic vesicle-associated membrane protein/synaptobrevin of tetanus and botulinum type B and G neurotoxins. *J. Biol. Chem.* **271**, 20353–20358 (1996).
- Cornille, F. *et al.* Cooperative exosite-dependent cleavage of synaptobrevin by tetanus toxin light chain. *J. Biol. Chem.* **272**, 3459–3464 (1997).
- Holden, H. M., Tronrud, D. E., Monzingo, A. F., Weaver, L. H. & Matthews, B. W. Slow- and fast-binding inhibitors of thermolysin display different modes of binding: crystallographic analysis of extended phosphonamide transition-state analogues. *Biochemistry* **26**, 8542–8553 (1987).
- Lacy, D. B., Tepp, W., Cohen, A. C., DasGupta, B. R. & Stevens, R. C. Crystal structure of botulinum neurotoxin type A and implications for toxicity. *Nature Struct. Biol.* **5**, 898–902 (1998).

- Hanson, M. A. & Stevens, R. C. Cocystal structure of synaptobrevin-II bound to botulinum neurotoxin type B at 2.0 Å resolution. *Nature Struct. Biol.* **7**, 687–692 (2000).
- Agarwal, R., Eswaramoorthy, S., Kumaran, D., Binz, T. & Swaminathan, S. Structural analysis of botulinum neurotoxin type E catalytic domain and its mutant glu212 → gln reveals the pivotal role of the glu212 carboxylate in the catalytic pathway. *Biochemistry* **43**, 6637–6644 (2004).
- Vaidyanathan, V. V. *et al.* Proteolysis of SNAP-25 isoforms by botulinum neurotoxin types A, C, and E: domains and amino acid residues controlling the formation of enzyme-substrate complexes and cleavage. *J. Neurochem.* **72**, 327–337 (1999).
- Foran, P., Shone, C. C. & Dolly, J. O. Differences in the protease activities of tetanus and botulinum B toxins revealed by the cleavage of vesicle-associated membrane protein and various sized fragments. *Biochemistry* **33**, 15365–15374 (1994).
- Washbourne, P., Pellizzari, R., Baldini, G., Wilson, M. C. & Montecucco, C. Botulinum neurotoxin types A and E require the SNARE motif in SNAP-25 for proteolysis. *FEBS Lett.* **418**, 1–5 (1997).
- Li, L., Binz, T., Niemann, H. & Singh, B. R. Probing the mechanistic role of glutamate residue in the zinc-binding motif of type A botulinum neurotoxin light chain. *Biochemistry* **39**, 2399–2405 (2000).
- Binz, T., Bade, S., Rummel, A., Kollewe, A. & Alves, J. Arg(362) and Tyr(365) of the botulinum neurotoxin type A light chain are involved in transition state stabilization. *Biochemistry* **41**, 1717–1723 (2002).
- Brooijmans, N., Sharp, K. A. & Kuntz, I. D. Stability of macromolecular complexes. *Proteins* **48**, 645–653 (2002).
- Sutton, R. B., Fasshauer, D., Jahn, R. & Brunger, A. T. Crystal structure of a SNARE complex involved in synaptic exocytosis at 2.4 Å resolution. *Nature* **395**, 347–353 (1998).
- Rupp, B. & Segelke, B. Questions about the structure of the botulinum neurotoxin B light chain in complex with a target peptide. *Nature Struct. Biol.* **8**, 663–664 (2001).
- Sukonpan, C. *et al.* Synthesis of substrates and inhibitors of botulinum neurotoxin type A metalloprotease. *J. Pept. Res.* **63**, 181–193 (2004).
- Schmidt, J. J. & Bostian, K. A. Proteolysis of synthetic peptides by type A botulinum neurotoxin. *J. Protein Chem.* **14**, 703–708 (1995).
- Fiebig, K. M., Rice, L. M., Pollock, E. & Brunger, A. T. Folding intermediates of SNARE complex assembly. *Nature Struct. Biol.* **6**, 117–123 (1999).
- Segelke, B., Knapp, M., Kadkhodayan, S., Balhorn, R. & Rupp, B. Crystal structure of Clostridium botulinum neurotoxin protease in a product-bound state: evidence for noncanonical zinc protease activity. *Proc. Natl Acad. Sci. USA* **101**, 6888–6893 (2004).
- Matthews, B. W. Structural basis of the action of thermolysin and related zinc peptidases. *Acc. Chem. Res.* **21**, 333–340 (1988).
- Dall'Acqua, W. & Carter, P. Substrate-assisted catalysis: molecular basis and biological significance. *Protein Sci.* **9**, 1–9 (2000).
- Fasshauer, D., Bruns, D., Shen, B., Jahn, R. & Brunger, A. T. A structural change occurs upon binding of syntaxin to SNAP-25. *J. Biol. Chem.* **272**, 4582–4590 (1997).
- Debnath, A. K., Radigan, L. & Jiang, S. Structure-based identification of small molecule antiviral compounds targeted to the gp41 core structure of the human immunodeficiency virus type 1. *J. Med. Chem.* **42**, 3203–3209 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank T. Binz and J. Ernst for providing initial BoNT/A and sn2 constructs and P. Adams, T. Fenn, S. Kaiser, Z. Panepucci, P. Strop and W. Weis for technical assistance and critical reading. Portions of this research were carried out at the Stanford Synchrotron Radiation Laboratory, a national user facility operated by Stanford University on behalf of the US Department of Energy (Office of Basic Energy Sciences). The SSRL Structural Molecular Biology Program is supported by the Department of Energy (Office of Biological and Environmental Research) and by the National Institutes of Health (National Center for Research Resources, Biomedical Technology Program) and the National Institute of General Medical Sciences. Portions of this research were conducted at the Advanced Light Source which is supported by the Office of Energy Research (Office of Basic Energy Sciences, Materials Sciences Division) of the US Department of Energy at Lawrence Berkeley National Laboratory. This work was supported in part by an NIH grant to A.T.B.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.T.B. (brunger@stanford.edu). The coordinates and structure factors have been deposited in the Protein Data Bank under the accession codes 1XTF (apo structure) and 1XTG (complex).

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:

Nevin Bayouni (4978)

UK/ RoW/ Ireland:

Andy Douglas (4975)

Nils Moeller (4953)

Irene Vigila-Alton (4944)

Scandinavia/ Spain/ Portugal:

Evelina Rubio Håkansson (4973)

Natureevents: Silke Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Advertising Production

Manager: Billie Franklin

To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Niamh Shields

European Satellite Office

Germany/ Austria/ Italy/

The Netherlands/ Belgium:

Patrick Phelan

Tel + 49 89 54 90 57 11

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

Sharon de Weert

Tel + 44 (0) 20 7843 4970

e-mail: s.deveert@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707

Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo

MG Ichigaya Building (5F),

19-1 Haraikatomachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami

e-mail: rasami@naturejpn.com

naturejobs

An equal chance

Could sexual equality be the hot new recruiting tool? After decades of under-representation, things are looking up for women. A few key institutions and programmes are turning problems, such as the 'glass ceiling' and taking time off to have children, into positive recruiting aids to attract the best young female scientists.

The glass ceiling meant that of the many women who entered science, few made it to professor and fewer still rose to the top ranks. But those who have broken through are paving the way for others to join them. The appointments of Shirley Tilghman as president of Princeton University and of Susan Hockfield as president of the Massachusetts Institute of Technology have served as a wake-up call in the United States. And in Sweden, the Karolinska Institute's new president, Harriet Wallberg-Henriksson, wants to create equal opportunities for women at her institute and in Singapore, where she holds a dual appointment (see www.nature.com/naturejobs/channels/graduate).

But women don't need to be in charge to generate opportunities for other women. Drug and biotechnology firms are creating more flexible working environments, with better daycare facilities, for example (see *Nature Rev. Drug Disc.* **3**, 981; 2004). And a number of organizations are offering 'restart' grants to help women get back into the scientific workforce after taking time off to have children (see *Nature Med.* **10**, 114–115; 2004).

Although such programmes don't cover the majority of opportunities for women, they do represent the vanguard, as they are sponsored by some of the world's leading organizations. Institutions that fail to create better environments for women to work in, and don't offer more opportunities for them to advance, will do so at their peril. They will be guilty of ignoring a huge pool of talent that others now have firmly in their sights.

Paul Smaglik

Naturejobs editor



WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS